Sensitivity of Semantic Signatures in Text Mining

Sri Ramya Peddada
West Virginia University

Sensitivity of Semantic Signatures in Text Mining

Sri Ramya Peddada

Thesis submitted to the

College of Engineering and Mineral Resources

at West Virginia University

in partial fulfillment of the requirements for the degree of

Master of Science

In

Electrical Engineering

Dr.Elaine M. Eschen, Ph.D., Chair

Dr.Alan V. Barnes, Ph.D.

Dr. Arun A. Ross, Ph.D.

Department of LCSEE

Morgantown, West Virginia

2010

Keywords: semantic signatures, text mining, information retrieval,

data clustering, document classification

Abstract

Sensitivity of Semantic Signatures in Text Mining

Sri Ramya Peddada

The rapid development of the Internet and the ability to store data relatively inexpensively has contributed to an information explosion that did not exist a few years ago. Just a few keystrokes on search engines on any given subject will provide more web pages than any time before. As the amount of data available to us is so overwhelming, the ability to extract relevant information from it remains a challenge.

Since 80 % of the available data stored world wide is text, we need advanced techniques to process this textual data and extract useful information. Text mining is one such process to address the information explosion problem that employs techniques such as natural language processing, information retrieval, machine learning algorithms and knowledge management. In text mining, the subjected text undergoes a transformation where essential attributes of the text are derived. The attributes that form interesting patterns are chosen and machine learning algorithms are used to find similar patterns in desired corpora. At the end, the resulting texts are evaluated and interpreted.

In this thesis we develop a new framework for the text mining process. An investigator chooses target content from training files, which is captured in *semantic signatures*. Semantic signatures characterize the target content derived from training files that we are looking for in testing files (whose content is unknown). The semantic signatures work as attributes to fetch and/or categorize the target content from a test corpus. A proof of concept software package, consisting of tools that aid an investigator in mining text data, is developed using Visual studio, C# and .NET framework.

Choosing keywords plays a major role in designing semantic signatures; careful selection of keywords leads to a more accurate analysis, especially in English, which is sensitive to semantics. It is interesting to note that when words appear in different contexts they carry a different meaning. We have incorporated stemming within the framework and its effectiveness is demonstrated using a large corpus. We have conducted experiments to demonstrate the sensitivity of semantic signatures to subtle content differences between closely related documents. These experiments show that the newly developed framework can identify subtle semantic differences substantially.

Acknowledgements

I would like to thank my advisors Dr. Elaine M Eschen & Dr. Alan V Barnes for their continued effort, guidance and inspiration during the course of this research. I also wish to thank Dr. Arun A. Ross for his valuable suggestions and support.

Special thanks to my student colleagues, Uday Kiran Para and Ravali Kota who helped me all through this journey. Finally, words alone cannot express the thanks I owe to my parents and Kranthi, my husband, for their immense faith and blessings. All this wouldn't have been possible without their support.

I would also like to thank the Lane Department of Computer Science and Electrical Engineering (LCSEE) at West Virginia University for giving me a chance to pursue my Masters Education.

Table of Contents

List of Figures

List of Tables

1: Introduction

The rapid technological advances in computers and networking technologies have made it easy to manage large amounts of data. The world's largest and fastest growing text database is the Internet. Large amounts of structured and unstructured data on the internet (World Wide Web) are stored in the form of WebPages, HTML/XML archives, E-mails and text files. Even in an organization, institution, company or on any local computer, the amount of information is overwhelming. The ability to access this information and transform it into knowledge, which can be useful in decision making in the corporate sector, is very crucial in the present world. Since the mid 1990s many researchers have been devising tools, techniques & methods that can be useful to organizations in identifying and extracting useful information

Do Prado et al. [1] has a opinion that in an environment where information is completely overloaded, concepts such as data, web and text mining have come in handy. These techniques borrow from other techniques such as artificial intelligence, statistics, databases and information retrieval aiming to scale them to the new problems. Text mining, in particular, has shown a considerable evolution from simple word processing to present day where the adequate processing of concepts or even the extraction of knowledge from linguistic structures has been made possible.

Indeed, there are numerous applications of text mining, including extensive research in the analysis and classification of news reports, emails filtering and spam reduction, topic extractions from web pages, automated information extraction and management. All these applications demand a perfect text corpora and a set of robust and highly scalable algorithms for the text analysis. A systematic framework for incorporating domain knowledge is essential for a successful application. Thus, the proposed algorithms should be flexible enough to learn the appropriate patterns in the text corpora and should include prior knowledge of the domain.

1.1 Motivation

Text mining has become extremely prevalent, giving rise to an age where vast amounts of textual information can be accessed, analyzed and processed in a fraction of second. The development of new technologies to tackle problems such as topic detection, tracking and trend detection is bound to have wide applications in the future.

Digital text data such as IRC/AOL chat messages, bulletin board postings, forums, web pages, emails, text files on seized disks can carry identifying patterns. These patterns can be used to identify/analyze content and identify individuals. In this project we are developing methods for quantifying content, intent, and emotive shift in text data. On these grounds, we were motivated to develop a framework for text mining with which the information extraction and retrieval will be possible.

Our basic approach to mining text data aims at capturing the semantic structures in the text. Semantic structure depends on the correlations between keywords and locality of keyword groups. The traditional bag-of-words or keyword frequency approaches fall short of modeling these attributes. Our approach models not only keyword frequency, but also the distance between keywords and their relative ordering in the text. To this end, we derive high-dimensional vectors that store quantified relationships between keywords in a text document. In order to capture the locality of semantic structures, we generate many vectors per document. The content of these vectors is similar to the document vector (one per document) used by Wu et al. in [2, 3]. However, unlike Wu et al., we do not use these vectors directly to classify documents. Vectors generated from known content (learning) documents are used to develop Semantic Signatures that model the semantic structure of the target content. Multiple Semantic Signature can be used to model various nuances of single target content. Semantic Signatures drawn from a library are then used to classify documents of unknown content. Our new approach has proven to be a remarkably sensitive tool for differentiating semantic content in text data.

1.2 Statement of the Thesis

This research includes the design and development of a framework for the text mining process. It includes the tools package called Semantic Signatures Mining Tool (SSMinT) which was developed using this framework. There are three tools in this package – Keyword Tool, Learner Tool and Data Analysis Tool. The methodology incorporated in the SSMinT package of tools are the design of keyword sets and development of semantic signatures (the fingerprints of the content in the document), which in turn act as the target content to capture the information of interest in a large corpus of data. We have also formulated three experiments to test whether SSMinT can capture the semantic subtle nature of the English language. We have conducted experiments to demonstrate the sensitivity of semantic signatures to detect the subtle content differences in closely related documents.

1.3 Structure of this Document

Chapter 2 gives a broad view of what text mining is about and its process. The general architecture and applications of text mining is discussed briefly. This chapter proceeds to a literature review on text mining and related work that can be quoted in the content of this thesis. This chapter concludes with the commercial and noncommercial text mining tools currently available in the market.

Chapter 3 brings an interesting approach on the design and development of the concept of the software package. This chapter gives a comprehensive review of the methodologies used in our text mining process. It gives the backbone framework in developing the package.

Chapter 4 describes how the prototype software works. This is the proof of concept of the proposed framework in the previous chapter. This chapter discusses a detailed overview about how each software tool functions.

Chapter 5 begins the experiments sections. Here the stemming concept is explored .we present an experiment to determine how effective stemming is when used with our semantic signature approach, in terms of document retrieval. A complete experiment is conducted with a large data corpus.

Chapter 6 presents two experiments on different types of corpora that investigate whether the tools identify the semantic subtlety of the language. The subtle differences in the content of two different domains are exposed to a set of training files, and we show that our tools can identify the difference in the concepts.

Chapter 7 extends to the conclusion and proposes future work in this area of research.

2: Background and Related Work

2.1 What is Text Mining?

According to Franke et al [4], Text Mining can be defined as a special case of data mining. Data mining deals with knowledge discovery in databases and is applied to numerical-structured data. Text Mining refers to the discovery of non-trivial, unknown useful information from large volumes of unstructured text files. Since its origin, Text mining is considered to be similar to data mining as the knowledge discovery in databases is applied to the text archives. Text mining is gaining lot of focus as 80% of the information (not considering other forms of media like audio, video etc) worldwide which is stored in computers consists of texts.

The rapid development of the World Wide Web has been tremendous. It is the fastest growing text database. The amount of data in an organization even on a local computer can be so overwhelming. Every employee undergoes a drill of searching relevant information in his organization at some point of time. Similarly, if a researcher has to get familiarize with his problem of interest, he need to read a vast number of academic papers. Text mining for similar reasons is gaining popularity as it can turn large databases of texts into new found information of interest which is valuable for a variety of purposes.

2.2 Text Mining Process and its Motivation

The objective of Text mining is the discovery of new unknown knowledge within the text collection or the text databases. Stavrianou et al. [5] has briefly explained the process of text mining.

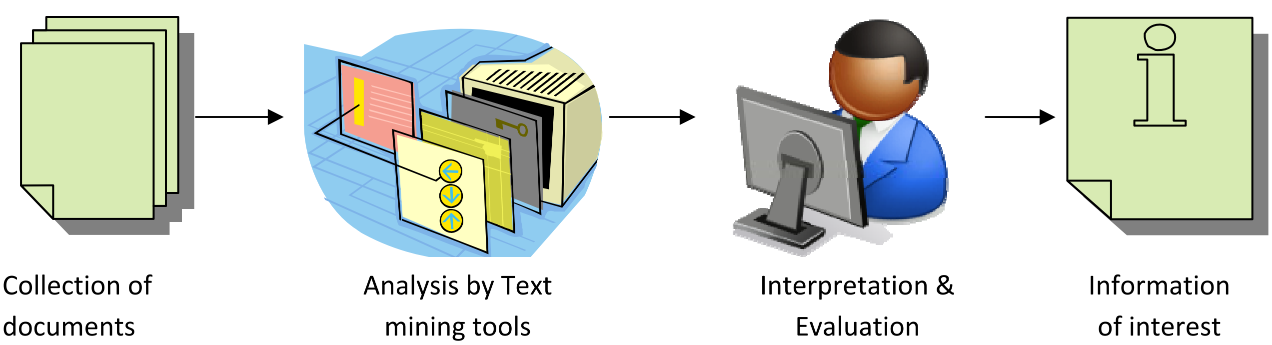


Figure 2.2.1: An Abstract Text Mining Process

The text mining process consists of data analysis of corpus/corpora as shown in the Figure 2.2.1. A text mining tool will perform data analysis on a collection of documents. During this process many sub-processes would take place like parsing, stemming, semantic and structural analysis, pattern recognition, clustering and tokenization. Following the analysis part, the interpretation of the tools output is needed. The results are evaluated and new found knowledge might emerge, which is the information of interest.

Data mining employs a number of machine learning algorithms which can also be extended to text mining. However, many issues arise with the limitations posed by natural language processing (NLP), which the aforementioned techniques do not always take in to consideration. An analyst needs to have a thorough understanding of the existing difficulties in text mining before he can work with them

The applications of text mining can extend to any sector where text documents exist. Stavrianou et al. [5] discussed many instances where text mining tools come to rescue. For instance, history and sociology researchers can benefit from the discovery of interesting patterns and links between events while crime detection agencies can benefit by the establishing similarities between one crime and another.

Text mining can definitely facilitate researchers. It can allow them to find related research issues related to the ones they are working on, retrieve references to past papers and articles which may have been forgotten and discover past methodologies that may augment recent research. Text mining has a capability to link two different research domains without putting an effort in understanding the texts within that domain.

Perhaps most notably, text mining exploits techniques and methodologies from the areas of information retrieval, information extraction and corpus-based computational linguistics.

Wikipedia has smartly listed several applications in each field [6].

Security applications: *ECHELON* surveillance system is one the leading and largest text mining applications available in the market. Many similar software packages like *AeroText, Attensity, SPSS and Expert System* have marketed towards security applications especially processing text sources such as text available in the internet.

Biomedical applications: A wide scope of text mining applications can be seen in biomedical literature. One such software to report is *PubGene* that combines biomedical text mining with network visualization and is available as an internet service.

Software and applications: IBM and Microsoft are some of the leading companies which are investing a lot of time and effort on text mining. They implement text mining techniques in the area of search and indexing in general as a way to improve their results.

Academic applications: The concept of text mining is of importance to publishers who hold large databases of information requiring indexing for retrieval. This is particularly true in scientific disciplines, in which highly specific information is often contained within written text. Therefore, initiatives have been taken such as *Nature's* (a popular scientific magazine) proposal for an *Open Text Mining Interface (OTMI)* and NIH's *common Journal Publishing Document Type Definition (DTD)* that would provide semantic cues to machines to answer specific queries contained within text without removing publisher barriers to public access.

2.3 General Architecture of Text Mining Systems

This section is reproduced in a brief manner from the book [1] "The Text Mining Handbook" by Feldman et al. [7] and is reprinted with the permission of Cambridge University Press.

At an abstract level, a text mining system accepts input (raw documents) and generates various types of output (e.g.: patterns, clusters, maps of concentrations, trends).On a functional level, text mining systems follow the general model provided by some classic data mining applications and are thus roughly divisible into four main areas **(a) preprocessing tasks (b) core mining operations (c) presentation layer components and browsing functionality and (d) refinement techniques.**

Preprocessing tasks include all those routines, processes and methods required to prepare data for a text mining system's core knowledge discovery operations. Preprocessing tasks generally convert the information from each original data source into a canonical format before applying various types of feature extraction methods.

Core Mining Operations are the heart of a text mining system and include pattern discovery, trend analysis and incremental knowledge discovery algorithms. Among the commonly used patterns for knowledge discovery in textual data are distributions (and proportions), frequent and near frequent concept sets, and associations.

Presentation Layer Components include GUI and pattern browsing functionality as well as access to the query language. Visualization tools and user-facing query editors and optimizers also fall under this architectural category.

Refinement Techniques at their simplest include the methods that filter redundant information and cluster closely related data .These involve comprehensive suite of suppression, ordering, pruning, generalization, and clustering approaches aimed at discovery optimization. These techniques have also been describes as post processing.

[1] © Ronen Feldman and James Sanger 2007

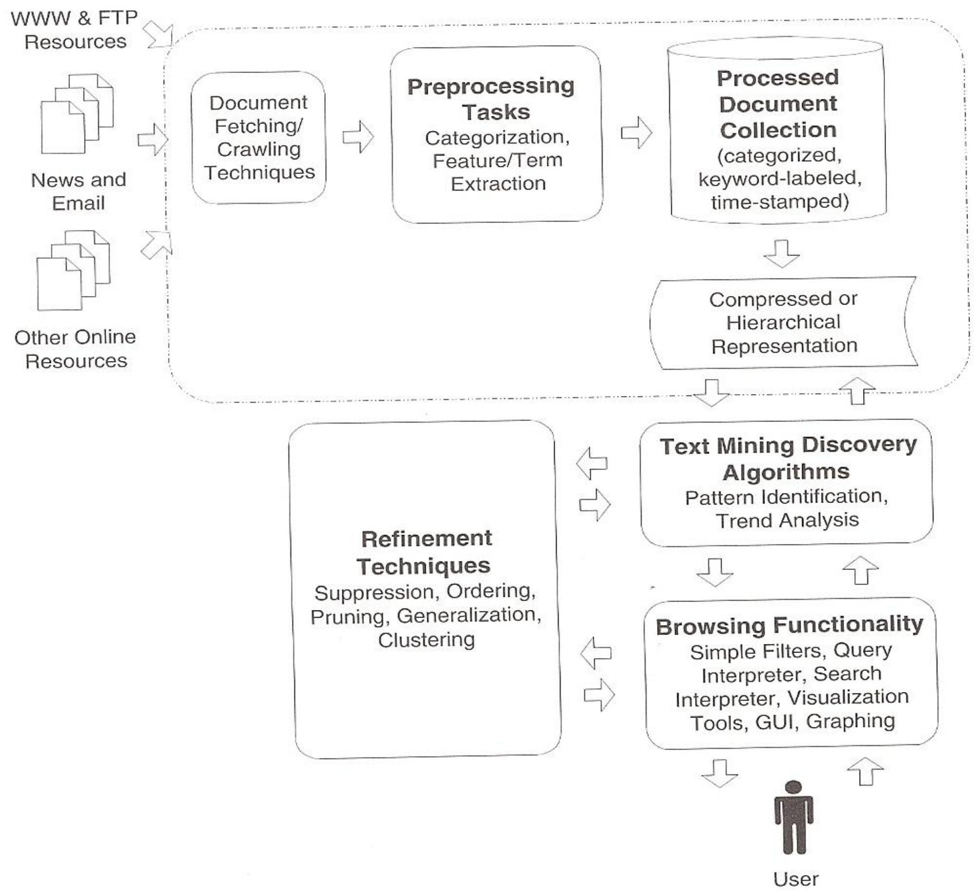


Figure 2.3.1: System architecture for generic text mining system[2]

At a slightly more granular level of detail, one will often find that processed document collection is, itself frequently intermediated with respect to core mining operations by some form of flat, compressed or hierarchical representation, or both, of its data to better support various core mining operations such as hierarchical tree browsing. This is shown in the System Architecture for generic Text mining systems.

[2] © Ronen Feldman and James Sanger 2007

2.4 Different Areas Where Text Mining is Used

Weiss et al. [8] has clearly listed several areas where text mining techniques are used. They are:

- Document Classification
- Information Retrieval
- Clustering and Organizing Documents
- Information Extraction
- Prediction and Evaluation

2.4.1 Document Classification

Text categorization or document classification means the same. It is the purest representation of the spreadsheet model with labeled destination results.

The below figure illustrates the document classification.

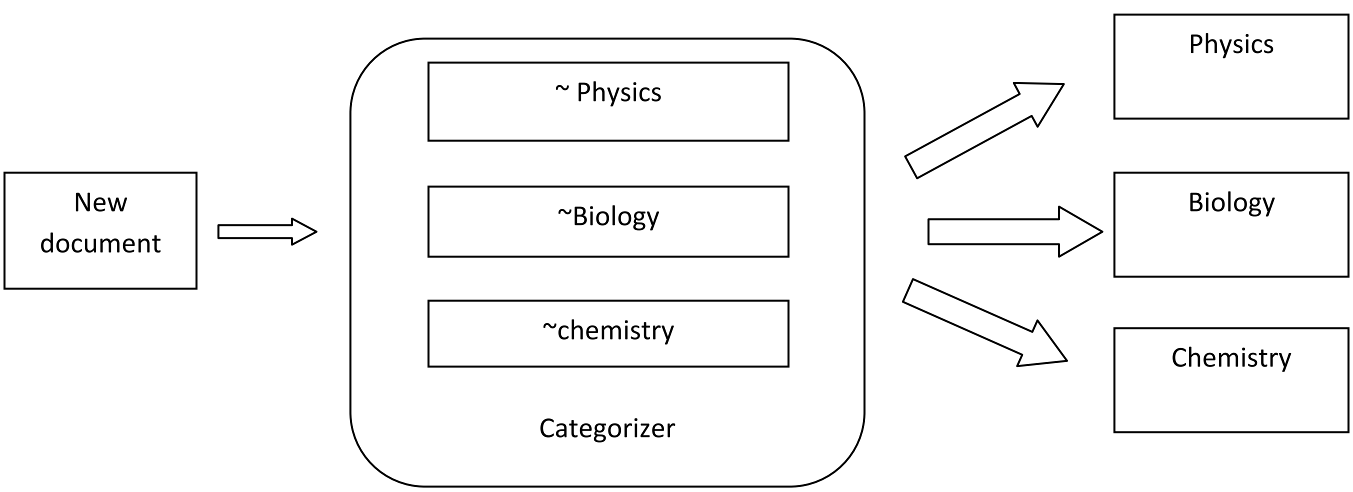


Figure 2.4.1.1: Text Categorization

Documents are organized into folders that represent each topic. When a new document is presented to the categorizer, its objective is to set that document to the appropriate folder. For example, we have a folder for physics, biology and chemistry research papers and we want to add new document to the correct folder. This classification is totally binary as the given document cannot be available in multiple folders. This type of categorization is called indexing, much like the index of a book. The adaptation of this task has broadened as more data has become available. For example, automatic email forwarding to the appropriate department is a type of text categorizer as it indexes to the email addresses available in that particular department.

2.4.2 Information Retrieval

Manning et al. [9] in their book has defined Information Retrieval as follows:

"Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from large collection (usually stored on computers)"

Abiding to its definition, information retrieval used to be an activity that only a few people engaged in: reference librarians, paralegals and similar professional searchers. Now according to the changes in the computing world and increase in information available on Internet, people are interested in information retrieval when they use a web search engine or search their email. Information retrieval has found wide range of applications and has overtaken traditional database style searching.

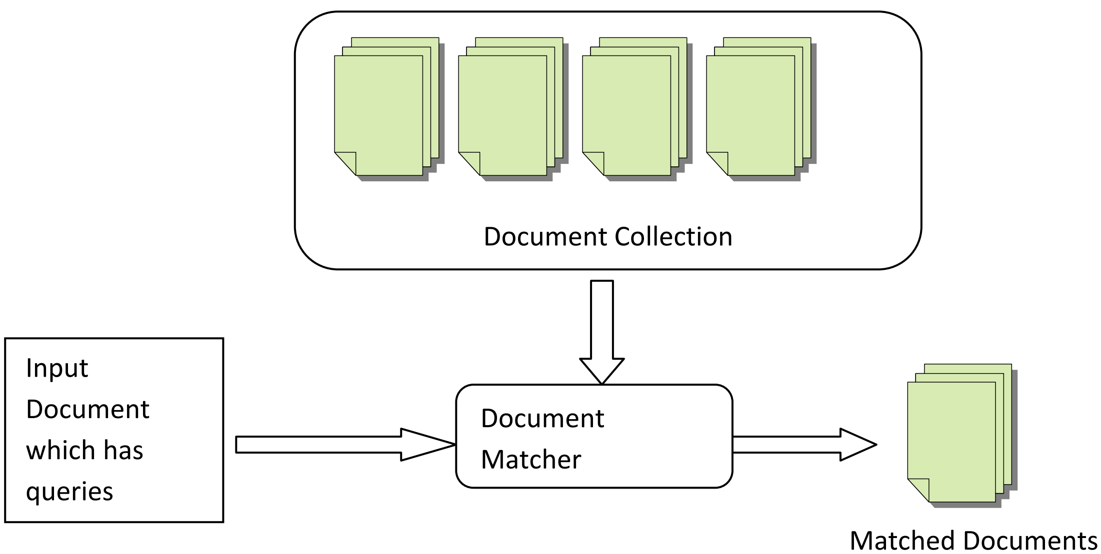


Figure 2.4.2.1: Document Retrieval

Basically, information retrieval is a search for similarity between two documents. In this technique even a small set of words to form a query and can help in retrieving the documents from a collection. From one perspective, measuring similarity is related to predictive methods for learning and classification that are called nearest-neighbor methods. In another perspective, IR is used to browse and filter the contents in a collection. The basic method of document retrieval is shown in Figure 2.4.2.1.

2.4.3 Clustering and Organizing Documents

Clustering is an unsupervised process through which objects are classified into groups called clusters. It groups similar objects into "more similar" fashion and dissimilar objects into a "more dissimilar" fashion. In categorization problems, as described in previous section, we are provided with a collection of preclassified training examples and the task of the system is to learn the descriptions of classes in order to be able to classify a new unlabeled object. In the case of clustering, the problem is to group the given unlabeled collection into meaningful clusters without any prior information. The labels associated with the clusters are again obtained by the input data.

Clustering is useful in a wide range of data analysis fields, including data mining, document retrieval, image segmentation, pattern recognition and text mining. In many such problems, little prior information is available about the data and the decision maker must make a few assumptions about the data if possible. It is for those cases the clustering methodology is especially appropriate.

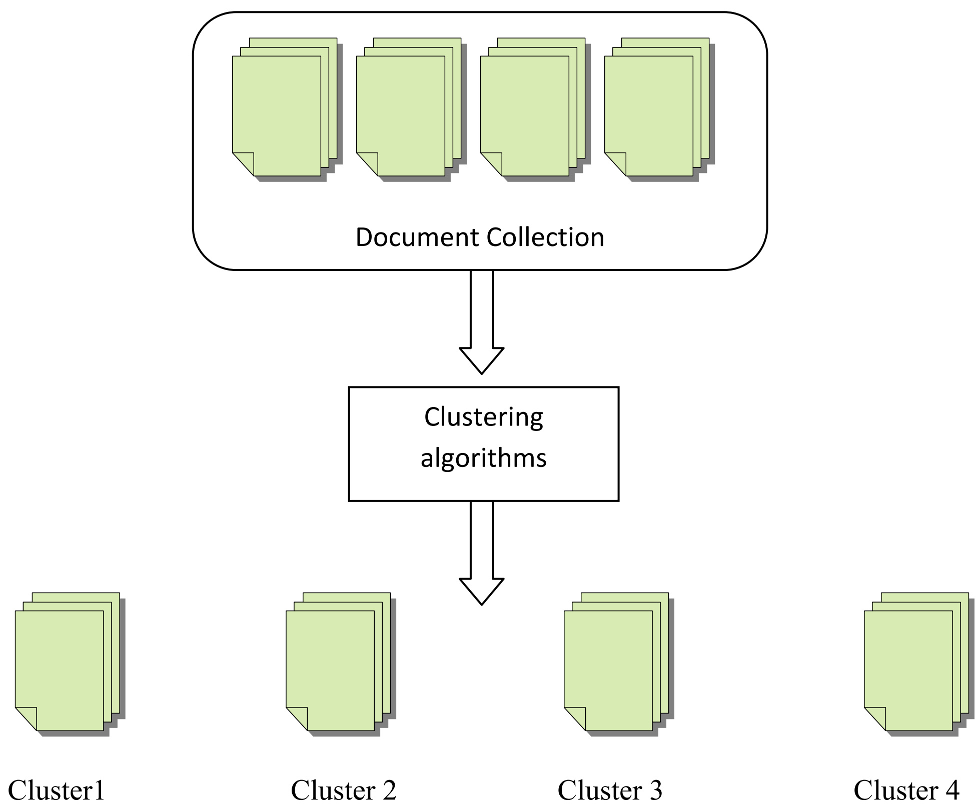


Figure 2.4.3.1: Organizing Documents in to clusters/groups

A clustering task may include the following components

- Define & represent the problem including feature extraction, selection or both
- Definition of proximity measure suitable to the domain
- Clustering the objects using algorithms
- Data augmentation, and
- Evaluation & interpretation.

2.4.4 Information Extraction

Information Extraction refers to the automatic extraction of structured information from unstructured sources. Structured information includes numbers, entities, relationship between entities, attributes describing the entities etc... This technique demands much richer queries when compared to keyword sets alone. When structured and unstructured data co-exist, information extraction makes it possible to link both the data and queries can be posed including both data. For over two decades, it has always been a challenging task for the researchers to extract information from a noisy unstructured source. Having its roots in the Natural Language Processing (NLP) field, the topic of structure extraction now engages many different communities including machine & statistical learning, information retrieval, database, web, and document analysis. Previously the extraction tasks include retrieving different entities from the text data like people, date and finding relationship between those entities. Now, Information extraction has also paved its way further and the scope of this research is so tremendous.

2.4.5 Prediction and Evaluation

Our ultimate goal is prediction, learn from the prior examples and project it to the unseen examples. The prediction algorithms learn by a learning program that studies the documents and finds some base to generalize a set of rules that will anticipate the correct results for new samples. But, how can we know whether the learning program was successful in predicting the new samples? The answer is to "hold out" some examples with known answers and not allowing the learning program to train on them. These new examples are used solely for evaluation. For many text-mining cases, the hold-out evaluation will be effective (e.g.: assigning labels to new brands, evaluating the scores). The challenge is, new samples change over time and we must keep track of its changes so that learning program is aware of them.

Measurement of error is one of the basic evaluations of the prediction technique. For evaluating scores, we can readily determine if the learning program has a "right" or a "wrong" prediction. The classical measures of accuracy will be applicable, but not all errors will be evaluated equally. Measurements of accuracy such as precision and recall are suitable for application in this domain.

2.5 Literature Review: Related Work

Text mining process includes text preprocessing, feature generation and selection, pattern extraction to analyze results. Many have contributed to the world of text mining and there are successful mining tools for both commercial and non-commercial purposes.

Amir et al. [10] describe a new tool called maximal associations that allows the discovering of interesting associations often lost by regular association rules. Hersh [11] evaluates different text-mining systems for information retrieval. Yang et al. [12] came up with a method of identifying the category theme automatically and hierarchical text categorization of Chinese language. Turmo et al. [13] introduce and compare different approaches to adaptive information extraction from textual documents and different machine language techniques. Saravanan et al. [14] discuss how to automatically clean data by discovering classes of similar items that can be grouped into prescribed domains. Srinivasan [15] develops an algorithm to generate interesting hypotheses from a set of text collections using Medline database. This is a fruitful path to ranking new terms representing novel relationships and making scientific discoveries by text mining. Van Heiist et al. [16] use data mining and boosting algorithms to create a support system for predicting end prices on eBay. Segall et al. [17] experimented with web text mining for hotel customer feedback using SAS® Text Miner and Megaputer Polyanalyst®

According to Wikipedia sources [6], there are approx 71 text mining tools available in the internet today. Many significant companies are investing their time and money to such new arena of research. Research and development departments of major companies, including IBM and Microsoft, are researching text mining techniques and developing programs to further automate the mining and analysis processes. Text mining software is also being researched by different companies working in the area of search and indexing in general as a way to improve their results

Text mining computer programs are available from a large number of commercial and open source companies. Below is the accepted list of applications listed in [6].

2.5.1 Commercial Software and Applications

- AeroText - provides a suite of text mining applications for content analysis. Content used can be in multiple languages.
- Attensity - hosted, integrated and stand-alone text mining (analytics) software that uses natural language processing technology to address collective intelligence in social media and forums; the voice of the customer in surveys and emails; customer relationship management; e-services; research and e-discovery; risk and compliance; and intelligence analysis.

- Autonomy - suite of text mining, clustering and categorization solutions for a variety of industries.
- Basis Technology - provides a suite of text analysis modules to identify language, enable search in more than 20 languages, extract entities, and efficiently search for and translate entities.
- Endeca Technologies - provides software to analyze and cluster unstructured text.
- Expert System S.p.A. - suite of semantic technologies and products for developers and knowledge managers.
- Fair Isaac - leading provider of decision management solutions powered by advanced analytics (includes text analytics).
- Inxight - provider of text analytics, search, and unstructured visualization technologies. (Inxight was bought by Business Objects that was bought by SAP AG in 2008).
- LanguageWare - text analysis libraries and customization tooling from IBM.
- LexisNexis - provider of business intelligence solutions based on extensive news and company information content set. Through the recent acquisition of Datops LexisNexis is leveraging its search and retrieval expertise to become a player in the text and data mining field.
- Mathematica provides built in tools for text alignment, pattern matching, clustering and semantic analysis.
- Nstein Technologies - text mining solution that creates rich metadata to allow publishers to increase page views, increase site stickiness, optimize SEO, automate tagging, improve search experience, increase editorial productivity, decrease operational publishing costs, increase online revenues. In combination with search engines it is used to create semantic search applications.
- SAS - solutions including SAS Text Miner and Teragram - commercial text analytics, natural language processing, and taxonomy software leveraged for Information Management.
- Silobreaker - provides text analytics, clustering, search and visualization technologies.
- SPSS - provider of SPSS Text Analysis for Surveys, Text Mining for Clementine, LexiQuest Mine and LexiQuest Categorize, commercial text analytics software that can be used in conjunction with SPSS Predictive Analytics Solutions. SPSS is now an IBM company.
- StatSoft - provides STATISTICA Text Miner as an optional extension to STATISTICA Data Miner, for Predictive Analytics Solutions.
- Thomson Data Analyzer - enables complex analysis on patent information, scientific publications and news.

2.5.2 Open-source Software and Applications

- GATE - natural language processing and language engineering tool.
- UIMA - UIMA (Unstructured Information Management Architecture) is a component framework for analysing unstructured content such as text, audio and video, originally developed by IBM.
- RapidMiner with its Text Processing Extension - data and text mining software.
- Carrot2 - text and search results clustering framework.

3: Design/Development of Concept of Software Package - SSMinT

The remarkable rate of progress in computing and networking technologies has made it very easy to collect and store large amounts of structured/ unstructured text data such as web pages, HTML archives, E-mails & other text files readily available for any end user. The users request maybe varied. He/she may not be interested in simply searching and retrieving a document, but rather want an overview of the document collection such as: what topics are covered, what kind of documents exist, are the documents somehow related and so on.

Given these requirements the user would not know what he/she is looking for. Therefore a data/text mining approach would be appropriate because, by definition, it is discovering interesting regularities or exceptions from the data, possibly without a precise focus.

To mine text, we need to first process it into a form that data-mining procedures can use. Out research goal is to come up with tools which analyze such processed data.

Thus the steps for handling the data, in our proposed approach can be recognized as:

1. Collecting Documents
2. Preprocessing the data in the documents – standardizing the text in the documents
3. Mapping the text data from words to clusters

3.1Collecting Documents

According to Weiss et al. [8], the first step in text mining is to collect the data that is relevant. Relevant documents may already be given or may be a part of the problem definition. For example, a webpage retrieval application may implicitly specify that the relevant documents are web pages. The next stage is to clean or standardize the data. Sometimes, the documents are collected from data warehousing which makes the task of cleaning the data easy, as they are already standardized.

In some applications, a data collection process like the web crawlers can be employed , which goes in to the World Wide Web and collects the documents of a given criteria.

Sometimes, the document sets are extremely large that we need some sampling techniques to manage them. For instance, a data stamp or a time stamp on the documents can be used as a criterion to sample for more relevant data. For research and development of text-mining techniques, more generic data is necessary. This is usually called a corpus. There are many corpora available today that are appropriate for some studies. As the importance of large text corpora became evident, a number of organizations took initiatives to coordinate activity and provide the distribution mechanism for corpora.

3.2 Preprocessing of Data

Data preprocessing is the name given to any type of processing that is performed on raw data in order to prepare it for another processing procedure. Data preprocessing is commonly used in the beginning in any data mining practice. Data preprocessing morphs the data into a format that will be more easily and effectively processed for the purpose of the user.

Why do we perform data preprocessing? Once several documents are collected, it is very common to find variety of documents in several different formats. Some documents may have been generated by a word processor with its own proprietary for mat. Others might have been generated by a simple text editor and saved as a normal text. Some may have been scanned and stored in the form of images. In order to read the textual data within the image; we need an image to text analyzer. Therefore, we see the need to standardize the text which is retrieved from the documents collected. Below is the flow of the levels in preprocessing which come handy in the type of textual data we have used for developing the tools.

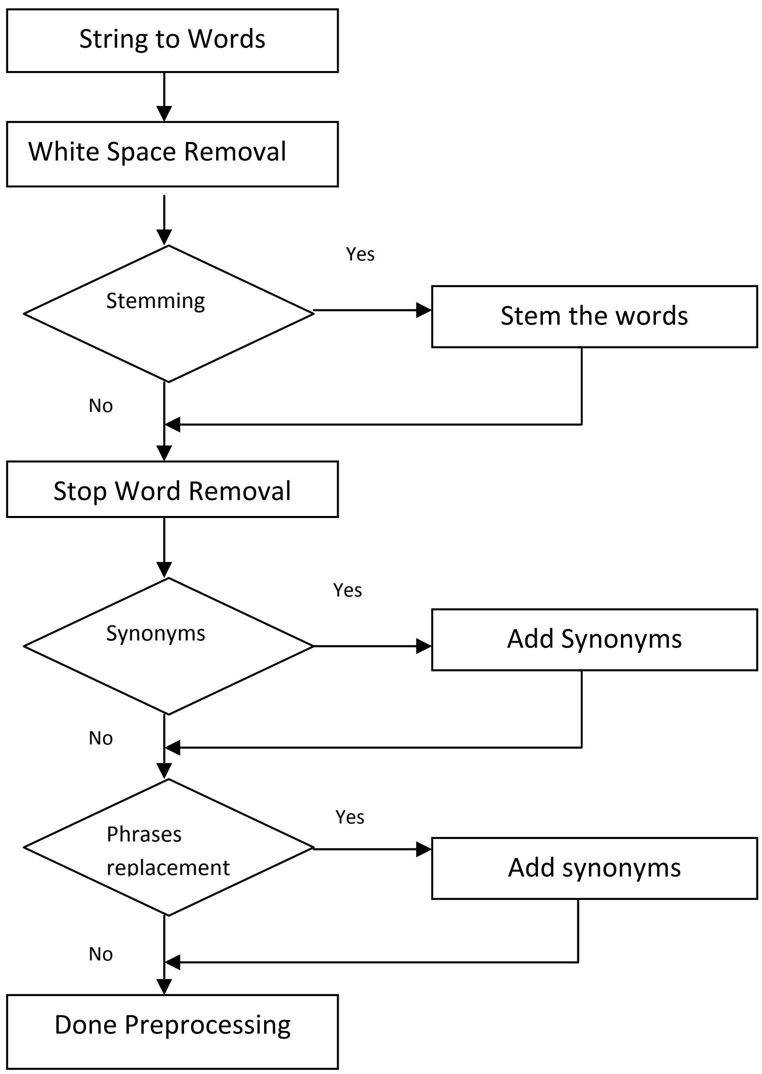


Figure 3.2.1: The flowchart representing the preprocessing/text handling in our proposed research design

3.2.1 String to Words

Preprocessing in our system starts with breaking the sentences in to words. Our methodology is keyword centric. To break the sentences in to words, we use string splitting, where the input file is read in to a string.

3.2.2 White Spaces Removal

According to computer science context, a single character or multiple character which represents horizontal or vertical space in typography is called whitespace character. A whitespace character does occupy the area in the page but doesn't leave a visual mark. For example, the common whitespace symbol " " (the Unicode character at the 32nd code point) represents a blank space, as used between words and sentences in Western scripts.

The term "whitespace" has originated from the idea that the background color to write any text is white. The most common whitespace characters may be typed via the space bar or the Tab key. Depending on context, a line-break generated by the Return key (Enter key) may be considered whitespace as well.

With respect to our text mining framework, once the words are converted in to strings, we removed all white spaces.

3.2.3 Stemming

Stemming is the process for reducing derived or inflected words to their stem, base or root form – generally a written word form. The stemmed word need not be identical to its root; it is sufficient that related words point to the same stem, even though the stem is not a valid root. The process of stemming is useful in web search engines for queries or information retrieval and other extraction problems. Stemming programs are commonly referred to as stemming algorithms or stemmers.

A stemmer for English, for example, should identify the string "stemmer", "stemming", "stemmed" as based on "stem". A stemming algorithm reduces the words "fishing", "fished", "fish", and "fisher" to the root word, "fish".

Based on Wikipedia sources [18], the first published stemmer was written by Julie Beth Lovins in 1968. This paper was outstanding and was a breakthrough of its age.

A later stemmer was written by Martin Porter and was published in the July 1980 issue of the journal Program. This stemmer gained its popularity and was widely used for English stemming. Dr. Porter received the *Tony Kent Strix* award in 2000 for his work on stemming and information retrieval.

Many implementations of the Porter stemming algorithm were written and freely distributed; however, many executions contained noticeable errors and as a result, these stemmers did not match their potential. To eliminate these different versions of errors, Martin Porter released an official free-software implementation of the algorithm around the year 2000. Over the years, he built a framework for rendering stemming algorithms called Snowball and implemented an improved English stemmer together with stemmers for several other languages.

The algorithm of porter stemmer is briefly explained in [19]. The Porter Stemmer is based on the idea that suffixes in English language are built with combining two or more suffixes. This stemmer is a linear step stemming algorithm. It has five steps applying rules within each step. Within each step, if a suffix matches the conditions within the step, the stemmable word undergoes suffix removal according to the rule defined within the condition and after removal, it moves to the another condition within the step. For example, if the number of vowels following a consonant is greater than one, then the vowel suffixes will be removed.

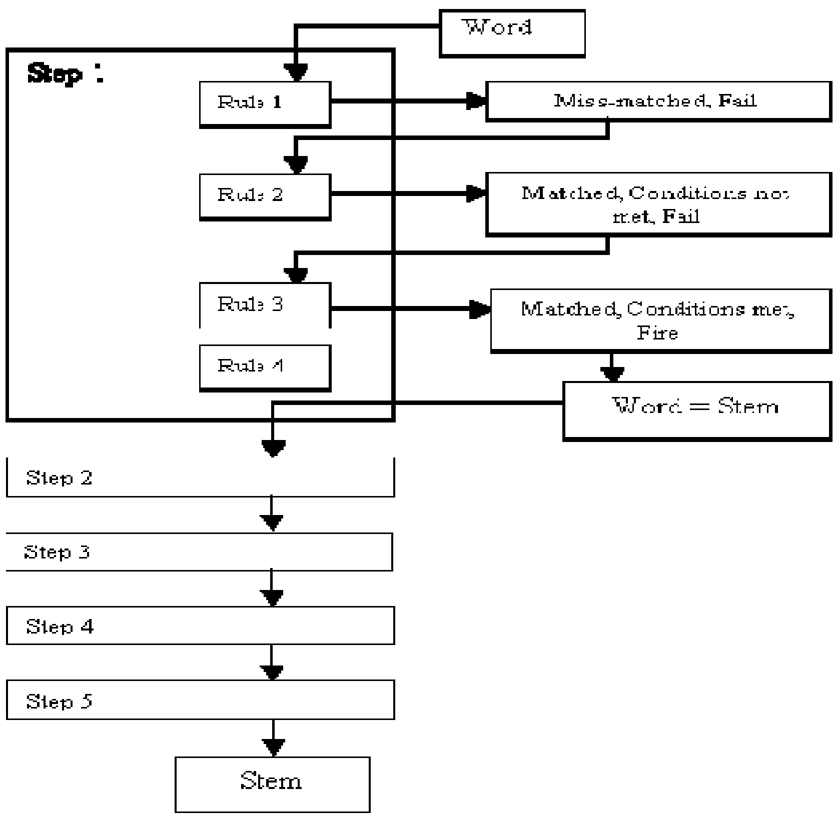


Figure 3.2.3.1: Porter Stemmer Algorithm[3]

If the rule is not accepted then the next rule in the step is applied and tested until either a rule from that step fires and control passes to the next condition or there are no more rules in that step when control moves to the next step. This process goes through all the five steps until every applicable rule is applied. The resultant stem being returned by the Stemmer after control has

[3] Courtesy of the figure is from Lancaster stemming algorithm website [14]

been passed from step five. See Figure 3.2.3.1. Due to its availability, Porter stemmer is widely used in many applications. Implementations of this stemmer are available at a website (*http://tartarus.org/~martin/PorterStemmer/*) established by Porter himself, with implementations in Java, C and PERL; the website also includes a copy of the paper defining the algorithm. Other implementations of this algorithm are available from the World Wide Web. Porter's algorithm is probably the stemmer most widely used in IR research.

We have integrated the Porter Stemmer in to our program. In the framework, Stemming is a part of preprocessor and is an option given to the analyst. It's the analyst choice if he wants to proceed to generation of keywords with or without stemming. This is an add-on feature in SSMinT package.

3.2.4 Stop Words Removal

Stop words is the name given to words which are filtered out prior to, or after, processing of text. As described in [20], Hans Peter Luhn, one of the originators in information retrieval, is credited with coining the phrase.

Stop words are less priority words which carry no meaning in the text. When it comes to queries in search engines, the stop words are removed from the query phase since they create more traffic. High stop word density can make any content look less significant.

Here, in our framework, we surely are going to face certain stop words which seem less important to our selection of keywords. Mostly we will end up with words that have high frequency to be 'a' or 'the' which are not of our interest. Thus, we adopted a stop words list (*http://armandbrahaj.blog.al/2009/04/14/list-of-english-stop-words/*) which we included in the keyword tool program to ignore the stop words. This list is referenced and if there is a stop word which appears to be in the high frequent words list, we remove such word.

3.2.5 Synonyms

Identical or similar meaning words are called synonyms. The words "develop" and "evolve" are synonyms. Similarly, if we talk about a long time or an extended time, long and extended become synonyms. In the figurative sense, two words are often said to be synonymous if they have the same connotation.

In our program, we created synonyms add-on feature in the generation of keywords. Synonyms can be stated in the program, and the Synonyms listed for a particular root word will be treated as the root word itself in the entire input file.

3.2.6 Phrases Replacement

Sometimes, depending u pon the typ e of input files, there is a necessity to not break the whole input file in to individual words. That is, two or more words together form a phrase that is m ore meaningful than splitting the phrase.

For example, consider the phrase “black market”. Black market is not a physical place, but rather an economic activity in which m erchandise and/or services are bought and sold illegally. So, if an input file contains infor mation talking about black m arket, we would not want the tool to break it in to black and market as two separate words. W e would want the sys tem consider this black market as a phrase. Phrase rep lacement methodology is included in the system as an add-on. W e can give phrases which w e want the sy stem to consider as a phrase. The system recalculates all the metrics according to the new input phrase.

3.3 Mapping the Textual Data

Suppose we have a corpus from which we want to ex tract text files with target content. W e want to pick certain keywords that are linked to the target content. Keywords com e in ha ndy as they represent the essential content of a document in condensed form. Keyword sets, which we define as a sequence of one or m ore words, provide a com pact rep resentation of targ et content. Keywords are widely used to define queries within text mining as they are easy to define, revise, remember, and share.
Arimura et al. [21] in his pa per, describes a fra mework which can discover the im portant keywords in the cyb erspace. We have an in tension to deve lop a fram ework for m ultiple uses of text mining like information retrieval, information extraction and prediction as well.

Figure 3.3.1 is a schematic overview of the design of our text mining framework.

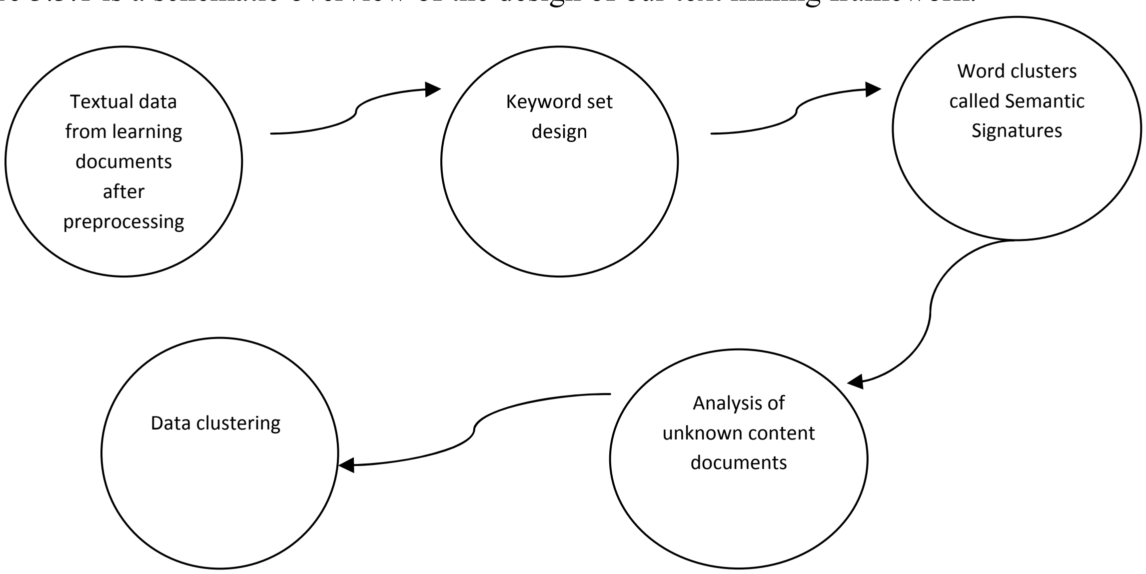


Figure 3.3.1: The process of mapping from words -> keyword sets -> Word Clusters

3.3.1 Keyword Set Design

Once the preprocessing is done, the words that were extracted from the file need to be processed to construct a successful keyword sets. Keywords here define the critical words that represent the content we would want to look for in future documents. Picking the keywords plays an important role in the design of an experim ent and is where the expertise of an analyst becom es valuable. If the keywords are correctly chosen, these can furthe r form a well-defined cluster of interest. This cluster of interest can be exposed to unknown documents to extract the necessary content.

After preprocessing, the learning document(s) are a sequ ence of words. The following are steps that are required to extract the keyword sets.

1. Words, which f orm the conten t of the f ile, are first ordered by fr equency. That is, high frequency w ords com e first. During this pr ocess we m ay ignore the stop words if they appear in the file (refer to stop words removal-preprocessing).
2. Once we have the word frequency list, the anal yst chooses a word that captures the target content, say KW0.
3. Window length confinem ent: The variable window length holds a certain num ber of words, for exam ple 20 or 50. Once the wi ndow length is fixed, the program accepts words in a window length for further processing.
4. Now, the program stores the first keyword KW0 and com putes a we ighted function that is a distance m etric within a window. This distance-m etric weighted function is defined as, w(x)

$$w(x) = \sqrt[2]{\frac{a^2}{x^2 + a^2}}$$

Where, 'a' is a constant defined by the user.

'x' is the keyword distance ; i.e., the distance measured by a word count between the keyword KW0 and the test word.

For e.g.: In the sentence, " *India is a country rich in her heritage. Our rich and colorful heritage, the soul of this great country, bestows on us our, own special identity, anywhere in the world.*"

5. Let's choose the KW 0 to be "country" and the test word to be "great". Now the word count between the two keywords is the keyword distance and is 6.
6. Keyword distance(in keyword selection only) always excludes the stop words like 'in', 'the', 'and' ,'her' ,'our', 'of', 'th is'. Thus, the weighted f unction between these two words is w (6) =0.409

"*India is a country rich in her heritage. Our rich and colorful heritage, the soul of this great country, bestows on us our own special identity, anywhere in the world.*"

7. These keyword distances are calculated forward and backward from KW0. For each of the words in the window, its weighted function is calculated from KW0, forward and backwards. This is maintained in a sorted list with largest weights first. That is, if a test word has more weight than another, it means that it is closer to KW0. Thus, the program gives this list of words to the analyst and asks the analyst to choose the next keyword KW1.

8. Once the KW1 is selected (the analyst should keep in mind the capturing of content and should pick keywords wisely) the program loops to the step 4 where now the active keyword is KW1 and this repeats until the analyst concludes the keyword set. Thus, two or more words form a set and the analyst can choose a successful keyword set with domain knowledge of the input content.

We now have a valid keyword set that is designed to capture the target content and we can begin further processing.

3.3.2 Semantic Signatures

To define a semantic signature, initially weighted vectors are identified in a given known content (learning) input text and these vectors are clustered with different clustering techniques. Selected clusters define Semantic Signatures. They are represented by the cluster centroid and radius of the cluster which holds the semantic vectors.

The three step process to define a semantic signature is

- Generating document vectors
- Clustering the document vectors
- Selecting a cluster (= semantic signature)

3.3.2.1 Generating Document Vectors

Semantic signatures are derived from a text file as clusters of document vectors extracted from that text file using keyword sets.

The explanation of the procedure of extracting semantic signatures is detailed with an example.

Consider the input text file contents to be:

"India is a country rich in her heritage. Our rich and colorful heritage, the soul of this great country, bestows on us our, own special identity, anywhere in the world. Arts & crafts are one very important aspect of our heritage. Each era has produced an art form unique to itself in expressing its beliefs and hopes. Thanjavur paintings of the Maratha period are a part of this rich art milieu. Today, this art is kept alive by a few hundred dedicated artists mostly based in Tamil Nadu - India.

Traditional yet Contemporary. Colorful to lift your spirits yet sublime to enhance spirituality. Divine in a prayer room, classy & elegant in other places. As gifts, unique and just beyond compare.

Welcome to the unique and colorful world of Thanjavur paintings. This school of paintings originated in Thanjavur during the reign of the Marathas in the 16th century. It existed from 17th to 19th Century, and had a limited output. Today, this tradition is kept alive by a few hundred dedicated artists mostly based in Tamil Nadu, India.

Thanjavur paintings basically signify paintings created using a style and technique, which originated in Thanjavur during the Maratha period in the 16th century. A typical Thanjavur painting would consist of one main figure, a deity, with a well-rounded body & almond shaped eyes. This figure would be housed in an enclosure created by means of an arch, curtains etc. The painting would be made by the gilded and gem-set technique - a technique where gold leaves & sparkling stones are used to highlight certain aspects of the painting like ornaments, dresses etc.

Traditional Thanjavur paintings are possessed as heirlooms. The painting would be bright & colorful and breathtakingly beautiful. The impact in a darkened room is that of a glowing presence. While most of the paintings would depict the Child Krishna and his various pranks, paintings of other deities were also created"

- Let's choose a set of three keywords in the file: India, Thanjavur, paintings. Let the window size be 20.
- The next step is to identify the active windows and the keywords.

"India is a country rich in her heritage. Our rich and colorful heritage, the soul of this great country, bestows on us our own special identity, anywhere in the world. Arts & crafts are one very important aspect of our heritage. Each era has produced an art form unique to itself in expressing its beliefs and hopes. Thanjavur paintings of the Maratha period are a part of this rich art milieu. Today, this art is kept alive by a few hundred dedicated artists mostly based in Tamil Nadu - India.

Traditional yet Contemporary. Colorful to lift your spirits yet sublime to enhance spirituality. Divine in a prayer room, classy & elegant in other places. As gifts, unique and just beyond compare.

Welcome to the unique and colorful world of Thanjavur paintings. This school of paintings originated in Thanjavur during the reign of the Marathas in the 16th century. It existed from 17th to 19th Century, and had a limited output. Today, this tradition is kept alive by a few hundred dedicated artists mostly based in Tamil Nadu, India.

Thanjavur paintings basically signify paintings created using a style and technique, which originated in Thanjavur during the Maratha period in the 16th century. A typical Thanjavur painting would consist of one main figure, a deity, with a well-rounded body & almond shaped eyes. This figure would be housed in an enclosure created by means of an arch, curtains etc. The painting would be made by the gilded and gem-set technique - a technique where gold leaves & sparkling stones are used to highlight certain aspects of the painting like ornaments, dresses etc.

Traditional Thanjavur paintings are possessed as heirlooms. The painting would be bright & colorful and breathtakingly beautiful. The impact in a darkened room is that of a glowing presence. While most of the paintings would depict the Child Krishna and his various pranks, paintings of other deities were also created"

- Windows are identified starting with a keyw ord. Only when there is another keyword appearance within the window length (here it is 20), it is said to be an active window.
- For the abo ve example, active wind ows have b een highlighted in grey and keywords in yellow, blue or green.
- In each window, the weighted functions be tween two keywords are calculated. For example, consider the active window

"……*India. Thanjavur paintings basically signify paintings created using a style and technique, which originated in Thanjavur during the Maratha period in*"…..

In this window, for each of the combination, India- Thanjavur-paintings, the weighted function is calculated.
Such combinations are
India-Thanjavur – 2 times
India-paintings – 2 times
Thanjavur –paintings – 1 time
Paintings- Thanjavur- 2 times
Thanjavur-thanjavur-1 time
Paintings-paintings-1 time

For the India and Thanjavur com bination, the word 'Thanjavur' appears imm ediately after India for the first tim e and after 15 words (including stop words) for the second tim e. Thus, the aggregate weighing function will be normalized between the two instances.

Weighing function for India-Thanjavur combination is

$\frac{w(1)+w(15)}{2}$=1.07901

Similarly, for other keyword combinations the weighted function is calculated as below:

'India-paintings' 1.3620
'Thanjavur-paintings' 0.9615
'Paintings-Thanjavur' 0.3288
'Thanjavur-Thanjavur' 0.113122
'Paintings-paintings' 0.7352

The above weighted function represented in a matrix form is:

	India	Thanjavur	Paintings
India	0	1.07901	1.3620
Thanjavur	0	0.11312	0.9615
Paintings	0	0.3288	0.7352

The 3 X 3 matrix is represented in a vector form as:

[0, 1.07901, 1.3620, 0, 0.11312, 0.9615, 0, 0.3288, 0.7352]

Once all the weighted functions are calculated in an active window and a vector is generated for the window, we move to the next active window. For a given input text file, a set of such vectors are generated.

3.3.2.2 Clustering the Document Vectors

Clustering the document vectors is essential to identify the vectors with similar orientation. Select vector clusters that capture the subject of interest, furthering our ability to identify the target content.

Clustering is a method that partitions a set of samples or observations (in this case, vectors) into subsets (or clusters) such that the members of a cluster are closely related in some sense. In other words, a cluster is a collection of objects that are "similar" to each other and are "dissimilar" to the objects belonging to other clusters.

We have initially chosen K-means clustering for the implementation as it is a very abstract level clustering, and is easy to implement.

K-means Algorithm

K-means algorithm described in Weiss et al. [8] is used to cluster the document vectors. K-means is one of the simplest unsupervised learning algorithms in clustering. The procedure follows a simple and easy way to classify a given document vectors in to K number of clusters or groups. The algorithms involves in defining K centroids, one for each cluster. These centroids happen to be random vectors in the given set. These centroids must be placed far from each other such that overlapping of the formed clusters is avoided. The next step is to take each remaining vector and group it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids. After we have these k new centroids, a new binding has to be done between the same set of vectors and the nearest new centroid. A loop has been generated and the process is repeated. As the process loops over we notice that the K centroid move their location for every time the new centroid is calculated until no more changes in the clusters happen. In other words centroids do not move any more.

The algorithm is composed of the following steps:

1. Place K points that represent the initial centroids into the space where the vectors to be clustered are defined.
2. Assign each vector to the group that has the nearest centroid.
3. Recalculate K new centroids when all the remaining vectors have been grouped.
4. Repeat Steps 2 and 3 until the centroids no longer move.

After a point, the procedure will always terminate, the k-means algorithm does not necessarily find the most optimal configuration. The algorithm is highly sensitive to the initial K centroids we choose. Solution to this randomization is to run K-means multiple times.

The clusters generated from the K-means algorithm are defined by the centroid and the radius of the cluster. We can choose clusters of vectors that capture the target content to be the Semantic Signatures.

3.3.3 Analysis of Unknown Content Documents

The objective behind this phase is to analyze a corpus of documents with unknown content along with the known content document(s) used as markers. The group of semantic signatures which were extracted in the previous phase embodies the target content and in this phase we will look for the same content in the corpus of unknown content documents.

The basic function of this phase is semantic feature detection; it detects the semantic features in the unknown content documents using the semantic signatures.

This detection is represented by a *semantic feature vector*. A matrix is corporate which has rows corresponding to semantic signatures. A row of the matrix forms a semantic feature vector for a corpus document. The matrix elements store the number of "hits" of a particular semantic signature by document vectors generated from a given corpus document.

For an unknown content document, the document vectors are generated .If the distance between these document vectors and the test semantic signature centroid is less than the radius of the test semantic signature, it is considered to be a "hit".

As a result this phase is totally automated. All it needs is the set of semantic signatures and a set of known and unknown content documents.

3.3.4 Data Clustering

The Document analysis matrix has row elements called s*emantic feature vectors.* These vectors whose elements indicate hits of the respective semantic signatures when subjected to clustering, helps us identify similar groups of unknown content documents sent in to the document analysis stage. We have known content documents in addition to the unknown content documents, known content documents acts as file markers to identify the genre of the cluster outcome.

Thus the Document Analysis Matrix which is the outcome of the document analysis stage is exposed to various clustering techniques .These techniques categorize the input unknown (+ known, if needed) content in to several clusters.

We rely on the third party tool called *Weka*, open source data mining software. Weka has several types of clustering techniques which will aid in interpreting document analysis matrix output.

4: Prototype Software: Proof of Concept

4.1 Introduction

The key objective of designing the tools is to find traces of certain target content in a pool of unknown data intelligently.

In the previous chapter we have studied the design and ideas behind the development of the software package. The SSMinT package was developed as a team in conjunction with Para [22] for his Master's Thesis dissertation.

We can clearly see that a set of three tools are required to get a complete solution. *Tool 1:* select the keyword set(s), *Tool 2:* develop semantic signature(s) *Tool 3:* search through the unknown content documents to find documents that have similar semantic content.

The tools in SSMinT work independently. The output of each tool is self-defined and can be used as an input to the next level of the tool. The outputs are represented in a uniform format so that it is easy for all the tools to be integrated. XML was chosen as the format to write the output from each tool. There are two reasons to do so. Firstly, XML offers a lot of functionality at a small cost. Secondly, XML can be well-understood and the output can be read by other programs with minimal effort.

In this chapter, we describe the tools in SSMinT. The tools in the package are:

Tool 1: Keyword Tool

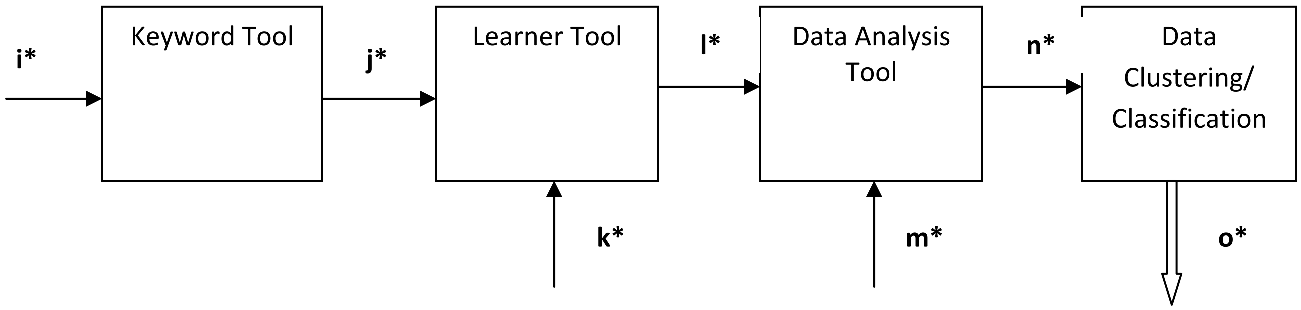
The main motive behind this tool is to select keyword sets that make sense semantically. These keyword sets are the backbone of the whole flow of the experiments. Once chosen wisely, keyword sets can make the tools efficient and robust in meeting the objective.

Tool 2: Learner Tool

The objective of this tool is to select the semantic signature, which are basically the cluster definitions that can capture the content revolving around the keyword sets chosen in the previous tool. Therefore, Learner Tool uses the output of the previous tool (keyword tool).

Tool 3: Data Analysis Tool

The Data Analysis Tool is designed keeping in mind the generalized search technique. This tool searches for some chosen content (embodied in the semantic signatures) in the unknown domain/content documents. Output is presented in a matrix/grid form which highlights the number of hits each semantic signature gets in each document of unknown content.



i* The known content files (training files) from which the keyword sets are chosen.

j* The output of the Keyword Tool called K*eyword Descriptor Files (KDFs)* - they define the keyword sets chosen in the keyword tool.

k* The training files (known content files used to generate the keyword sets).

l* The output of Learner Tool called, *Semantic Signature Descriptors (SSDs)* - they define the clusters that capture the target content.

m* The corpus of data with unknown content documents and known content documents included as markers.

n* *Document Analysis Matrix* generated from the Data Analysis Tool.

o* Final clustered/classified output.

Figure 4.1.1: An overview of the flow of information between the tools in the SSMinT software package

4.2 Keyword Tool

The motive behind developing this tool is to provide a user-friendly interface for choosing the right keywords that play a key role in identifying the desired content. Keyword Tool is developed such that it has different preprocessing techniques available to aid the selection of appropriate keywords. Different preprocessing techniques, which were discussed in the previous chapter, are employed in Keyword Tool to make it more robust. Also, an important feature is the point back feature that proves to be very useful in choosing the right keywords for a given content.

Below is the screenshot of Keyword Tool:

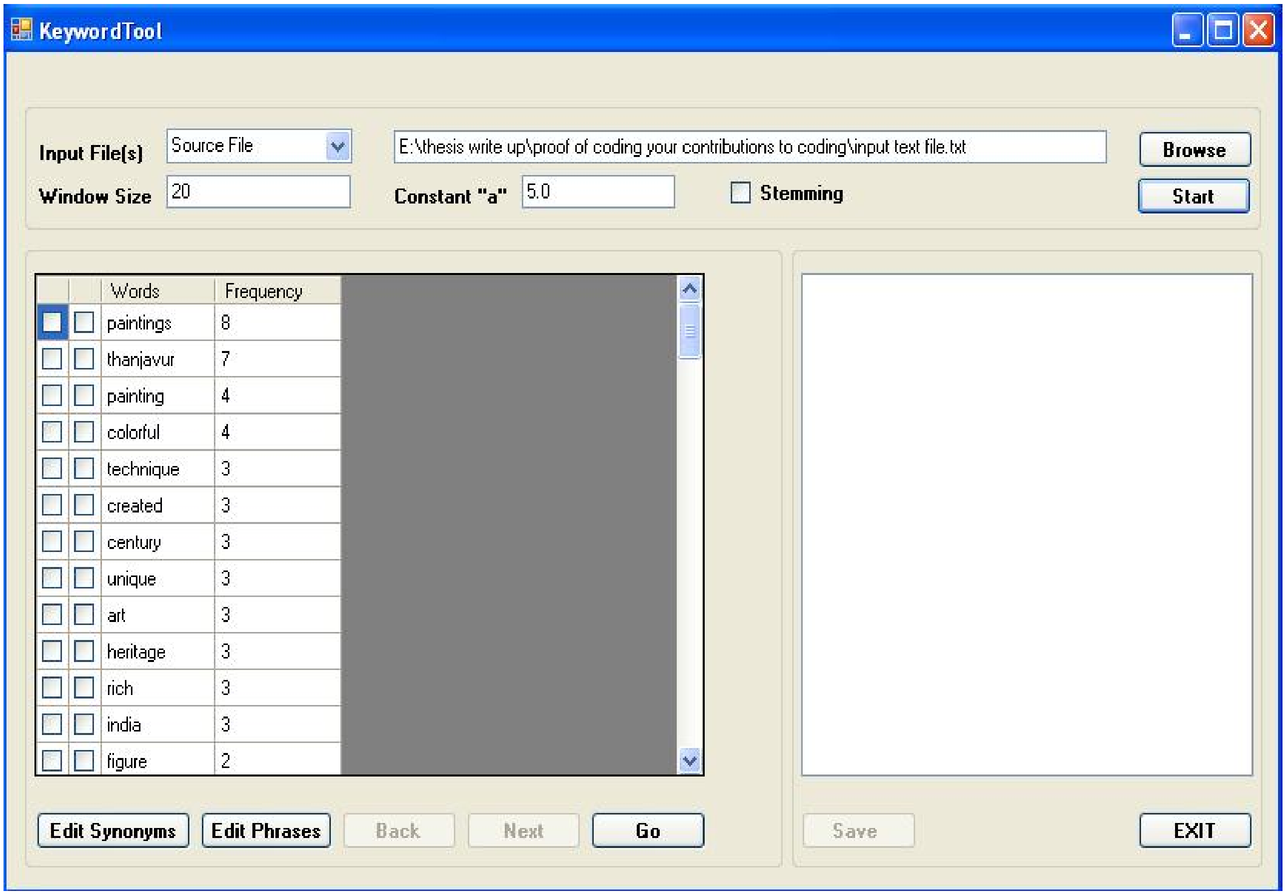


Figure 4.2.1: Screenshot of the Keyword Tool GUI

The user interface of Keyword Tool is shown in Figure 4.2.1. The process of choosing the keywords starts with loading a known content file into the tool. In the top right corner of the Tool's GUI there is a **Browse** button that when clicked opens a dialog box where the desired file location can be browsed and loaded to the tool. Once the file is loaded the **Window size** (as defined in the previous chapter, it is the maximum number of words in a window) is selected. Its default value is 20; i.e., 20 words are treated as a window in the program. The **constant** in the equation of the weighted function w(x) takes a user defined numeric value which defaults to 5.

$$w(x) = \sqrt[2]{\frac{a^2}{x^2 + a^2}}$$

As mentioned in the previous section, the robustness of the keyword tool is strengthened by the preprocessing techniques that are incorporated in the tool. Techniques like **stemming, phrase replacement (mega-words), synonym substitution and point back to text sources** aid in choosing the right keyword sets.

Stemming is an option in the GUI (located in the top portion of the GUI). If checked, all the words in the input file are stemmed to their root stem word. For example: the words "fishing", "fished", "fish", and "fisher" are stemmed to the root word, "fish". Here, we have integrated the **Porter stemming algorithm** as a plug-in to the keyword tool. The algorithm is very concise (having just about 60 rules) and readable for a programmer. It is also very efficient in terms of computational complexity. The main flaws and errors (for example; over-stemming for “police/policy”) are well known and can be corrected to an extent with a dictionary.

After choosing the input file and setting the input parameters like window size, constant and stemming, the **Start** button is initiated. Initially, the tool scans the whole file and displays the top 100 high frequency words in the data grid available in the left lower part of the GUI. As, you can see in Figure 4.2.1, the top 100 frequent words have been listed with their respective frequency measure.

In the data grid view, the first column of checked boxes is used for **point back to text sources.** This is also a plug-in to Keyword Tool. We can “point back” to the source file to see the whole source file in which the checked word is highlighted.

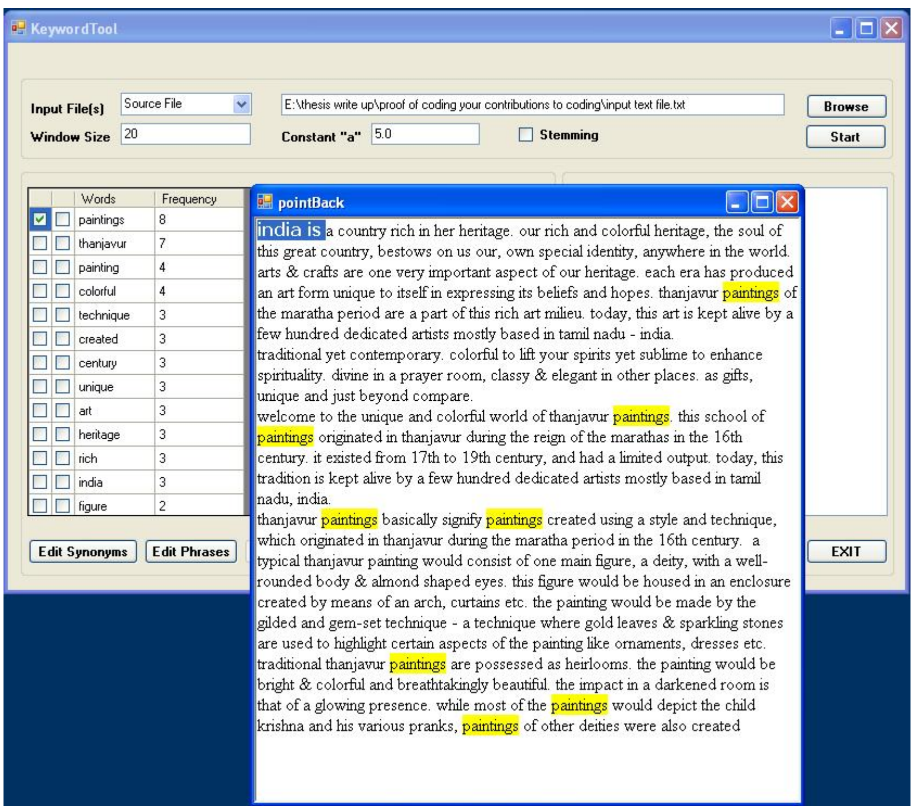


Figure 4.2.2: Point back window highlighting the checked word ‘paintings’

Let us choose the first keyword to be ***India.*** For this to happen, we have to use the second column of the check boxes in the data grid. Choose the check box against ***India*** and click on the button **Go** (which is located at the bottom right corner of the data grid). The right side lower part of the GUI has a list that stores selected keywords. The data grid refreshes and populates with the words that are close with respect to the weight function w(x) to the first keyword **India.**

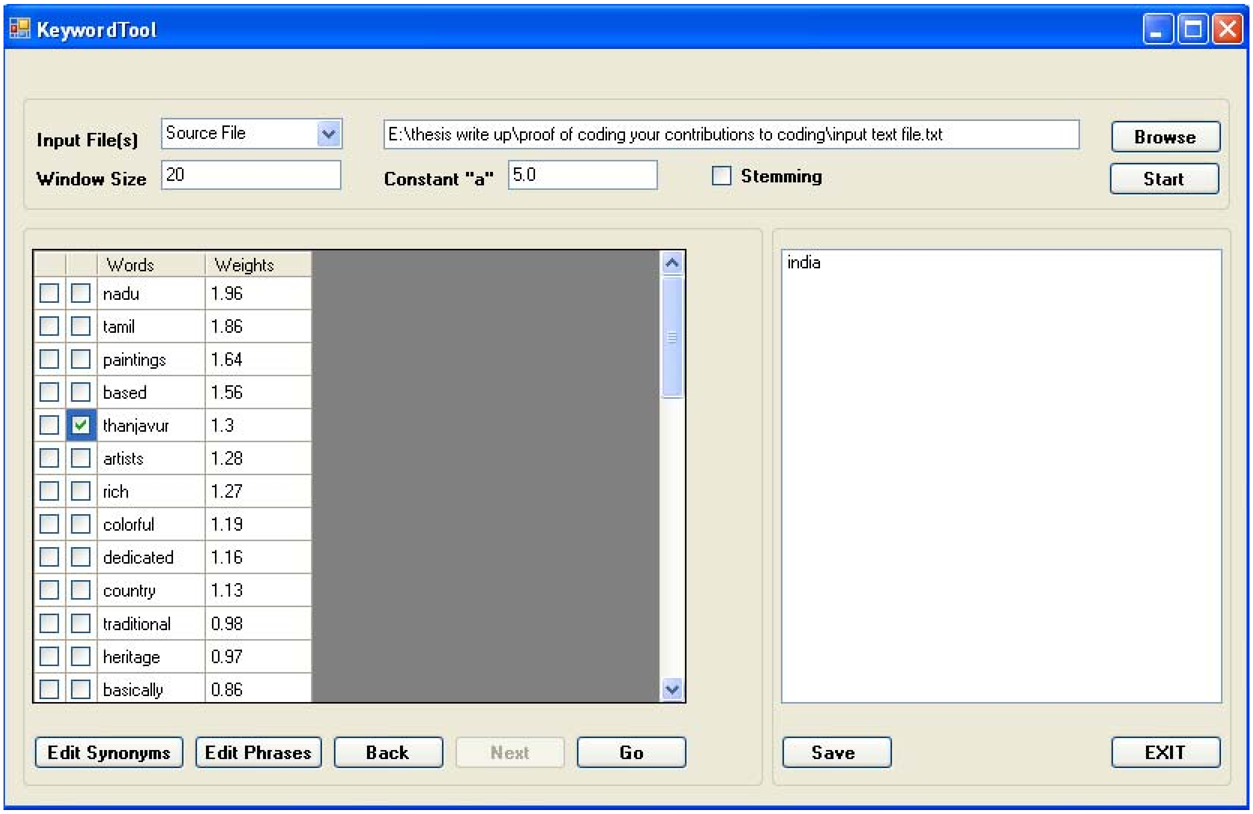


Figure 4.2.3: Data grid populating the words nearer to India

Let the second keyword be ***Thanjavur.*** Selecting ***Thanjavur*** and proceeding further, the data grid displays a list of words that are close to ***Thanjavur***.

Observe the **Back** button which is user-friendly. It lets the tool undo the last keyword selection and refreshes the data grid with the list of words that are nearer to the last keyword in the righ t-hand side list.

Edit Synonyms is another preprocessing plug-in in the tools, which lets the tool add synonym s for the words of interest.

The synonyms that are listed will be treated the sa me as the keyword/root word. In the program , the synonyms will be substituted by the root word. Synonyms are separated by commas (,). Also, the synonym 's GUI is user-friendly in adding/d eleting the root word & synonym pair at any point in the process.

Synonym's GUI has an **Add** button click, which accepts the text entered in the keywords and synonyms textboxes. After adding the synonyms we can click the **Finish** button. This again refreshes the data grid view with the changes after incorporating the synonyms change.

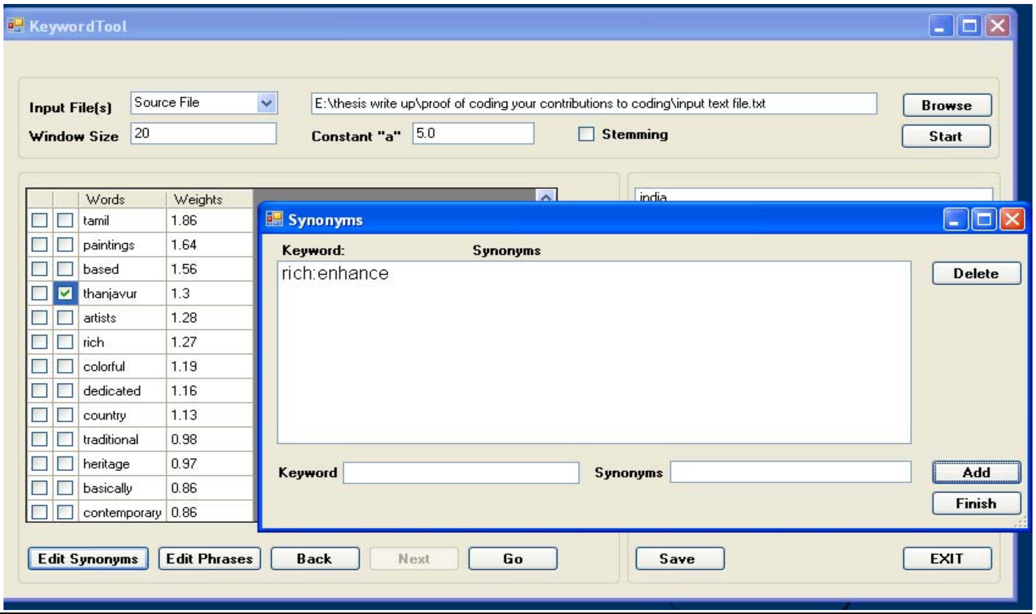


Figure 4.2.4: Adding synonyms GUI

Similarly, **Edit Phrases** is another plug-in which is a preprocessing technique. Adding a phrase to the input file is simply letting a phrase (group of words) is treated as a whole word entry. Similar to **Edit Synonyms** plug-in, the **Add, Finish, Delete** buttons work for the same purpose.

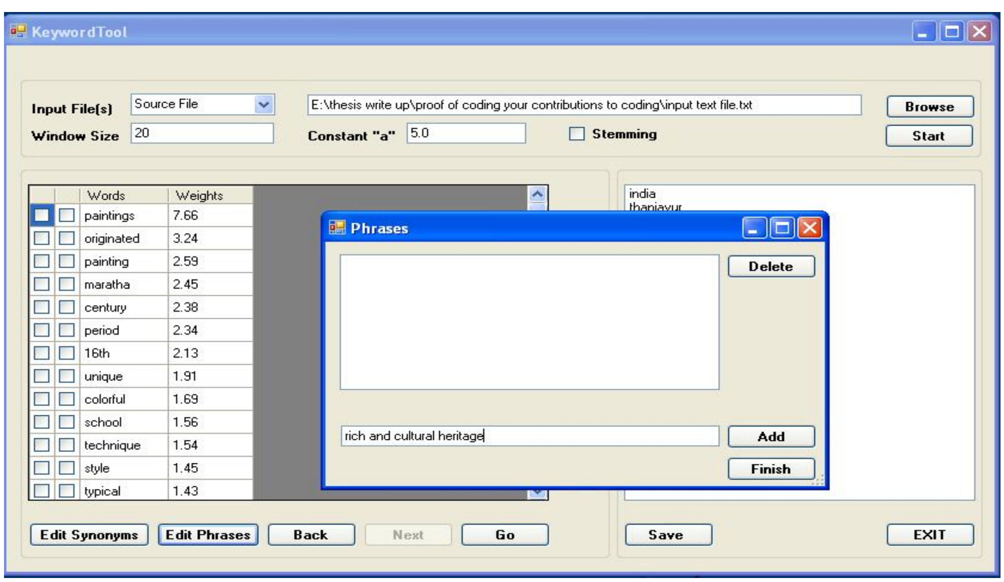


Figure 4.2.4: Adding phrases GUI

Phrases also can be added at any point of time in Keyword Tool.

Moving forward with the example and finishing the keyword selections with choosing keyword ***paintings*.** Below is the screen shot of the final keyword selection and the **point back source.**

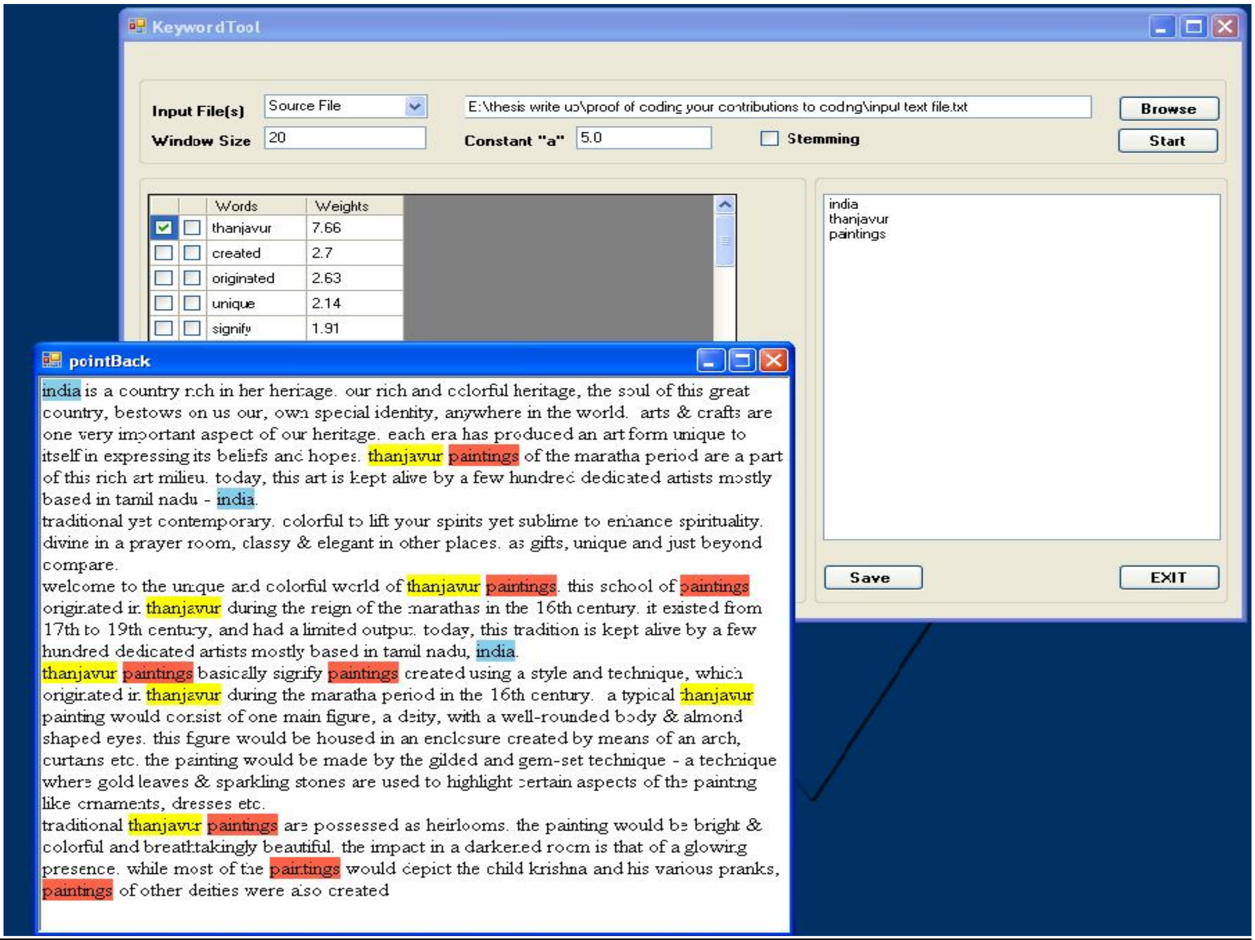


Figure 4.2.5: Keyword Tool after selection of all the keywords and showing point back source

4.2.1 Keyword Descriptor File - Output of Keyword Tool

The lower right hand side of the Keyword Tool GUI has a **Save** button. This lets us save the selected keywords in XML format. As quoted in the introduction section, this data format can be extended and applied broadly in several domains. The output of Keyword Tool is called the **Keyword Descriptor File (KDF).** The output of Keyword Tool is written out using a stream writer to a .kdf format file.

A KDF file looks like this:

```
<keywordTool version="1.1">

<stemming used="no" stemmer="porter"></stemming>

<source folder="no" url="no" file="yes">E:\thesis write up\proof of coding your
contributions to coding\input text file.txt</source>

<windowLength length="20"></windowLength>

<keywords>india,thanjavur,paintings

</keywords><synonyms></synonyms>

<phrases></phrases>

</keywordTool>
```

Figure 4.2.2.1: A sample .kdf file

The .kdf file starts with a *keyword version*: this is intended for the future, if the format of the KDF file changes, the version number can track the changes.

Stemming: indicates whether stemming is used or not and also lists the type of stemmer.

Source: indexes the file/folder location and indicates if a file or folder of files is used.

Window length: stores the window length used in the tool.

Keywords: lists the keywords selected in the tool.

Synonyms & phrases: lists synonyms and phrases, if used.

The .kdf file is the input to the learner tool, as the learner tool employs an xml reader to read the .kdf file and extract the necessary information.

4.3 Learner Tool

Learner tool generates semantic signatures. It operates on (a) the known file content which is used to generate the KDF and (b) the KDF itself. Let us take a closer look at developing semantic signatures with the help of the Learner tool

Below is the screen shot of the learner tool.

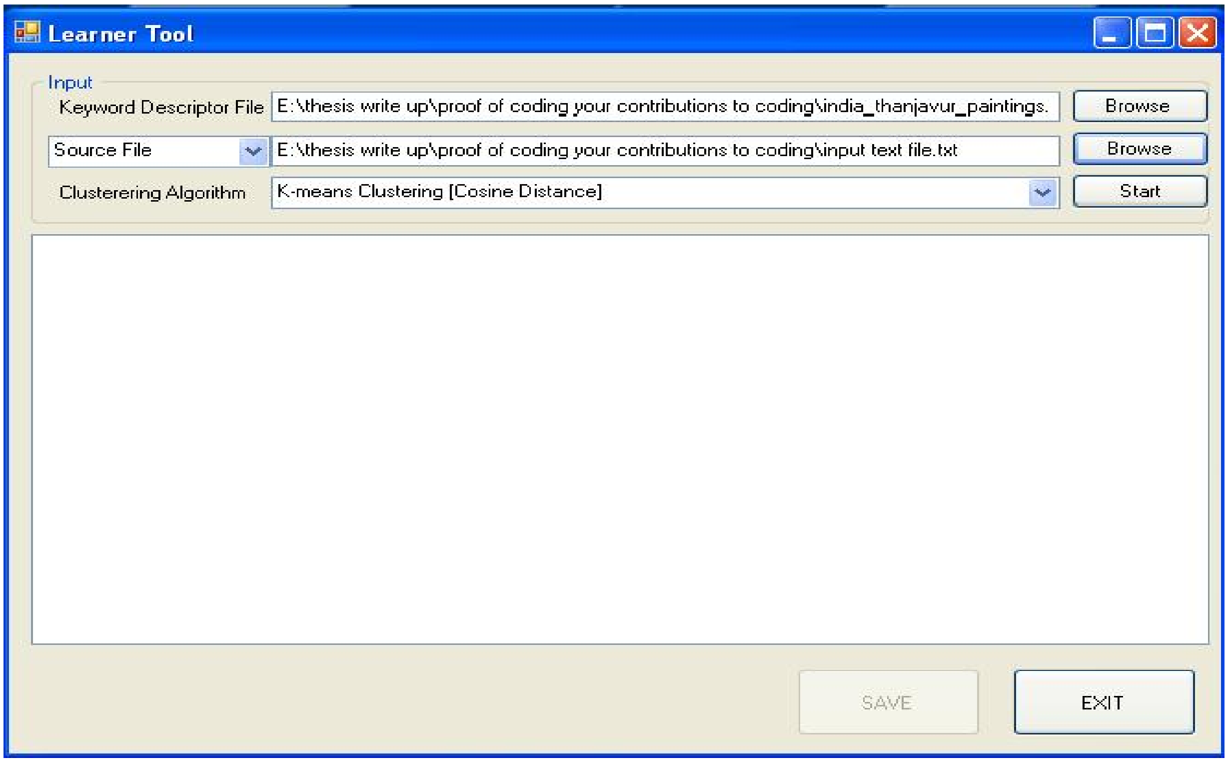


Figure 4.3.1: Learner Tool GUI

As you can see in the screenshot, the KDF file and the source file is indexed to its location by using a browse button (open file dialog box).

In the previous chapter we have stated that there are two steps in generating a semantic signature:

- Generating the document vectors

- Clustering the document vectors

For the clustering, we can select in the GUI the type of clustering from the drop down box in the GUI.

Once the clustering type is selected (before hitting start) the GUI asks for the number of clusters, if appropriate, in another dialog box. Once the number of clusters is entered, the program computes the document vectors.

Once, the **Start** button is clicked, the program first extracts the KDF information such as keywords and window length.

Throughout the source file, the document vectors are calculated. The document vectors are clustered using the clustering technique selected from the GUI. For now, we have introduced a simple clustering algorithm, K-means algorithm implemented to use either Euclidean or cosine distance measures.

The clustering is done on the vectors and is displayed in a tree view which provides point back to the original text so that the analyst can identify classes/clusters of vectors that embody the targeted semantic content.

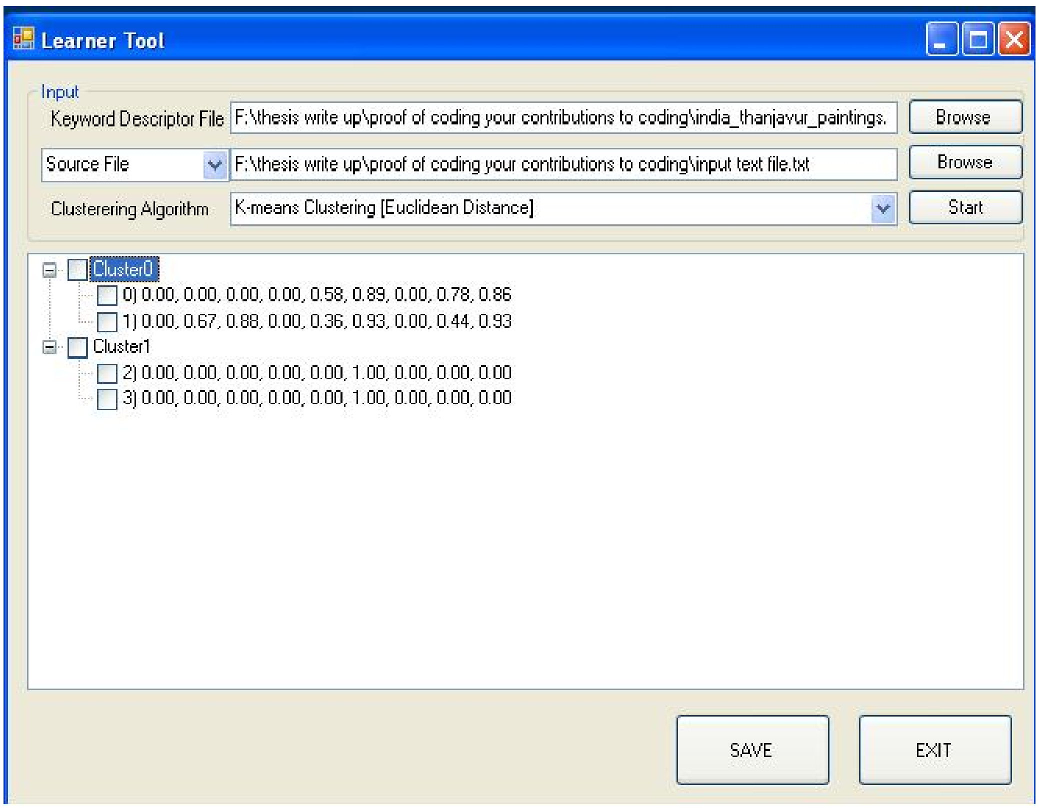


Figure 4.3.2: Learner Tool displaying the clusters

The clusters are displayed in the tree view where against each document vector there is a check box. When checked, it points back to the window of the source file text from which the vector was derived.

When a cluster is selected, this contains the document vectors of the target content. This cluster can be defined as a **Semantic Signature. Semantic Signature Descriptors (SSDs)** are the output of Learner Tool. The **Save** button saves the SSD in an XML format in a .ssd extension file.

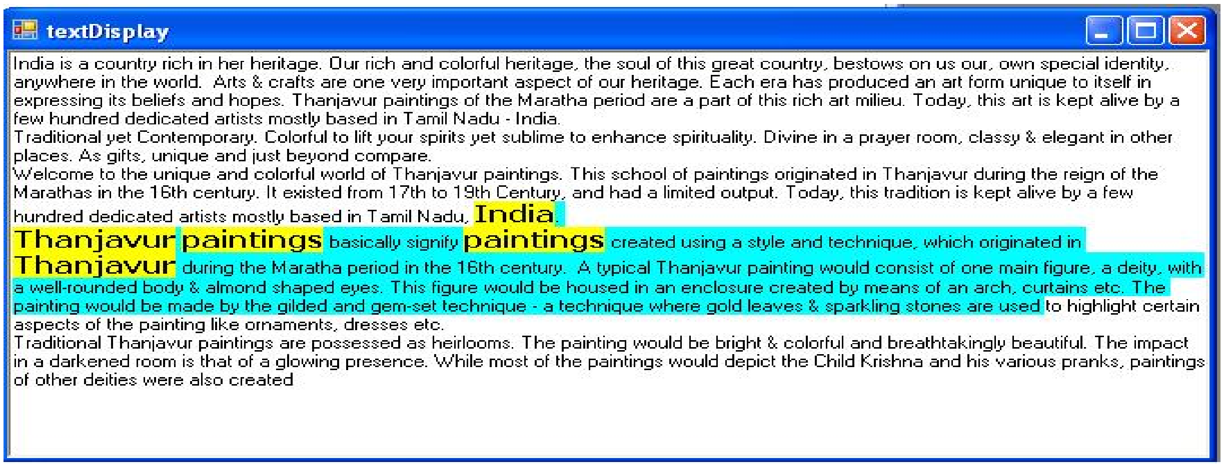


Figure 4.3.3: Point back text –Learner Tool

4.3.1 Semantic Signature Descriptor-SSD

Here is a sample SSD file


```
<ClassificationTool version="1.1">

<kdfSource>F:\thesis write up\proof of coding your contributions to coding\india_thanjavur_paintings.KDF</kdfSource>

<source folder="no" file="yes">F:\thesis write up\proof of coding your contributions to coding\input text file.txt</source>

<clusterer name="kmeans">2</clusterer>

<centroid r="0.591602071762943" distanceMeasure="ED">0, 0.3341, 0.4404, 0, 0.4701, 0.9096, 0, 0.6097, 0.893</centroid>

<vectors>0,0,0,0,0.5812,0.8904,0,0.7843,0.8575;0,0.6682,0.8807,0,0.359,0.9287,0,0.4351,0.9285;</vectors>

<keywordTool version="1.1">

<stemming used="False" stemmer="porter"></stemming>

<source folder="no" url="no" file="yes">E:\thesis write up\proof of coding your contributions to coding\input text file.txt</source>

<windowLength length="20"></windowLength>

<keywords>india,thanjavur,paintings

</keywords>

<synonyms></synonyms>

<phrases></phrases>

</keywordTool>
```


Figure 4.3.1.1: Learner tool output file: SSD

Similar to the KDF format, the SSD format is also in XML format. The KDF information is concatenated to the SSD. The centroid and radius of the SSD are defined. The type of clustering, distance measure and source files are also listed. Along with the centroid and radius, the vectors in the clusters are also stored in the SSD.

4.4 Data Analysis Tool (DAT)

Data Analysis Tool is the only tool in the SSMinT package that needs minimum analyst input, as this tool is fully automated. DAT operates on a corpus of data (plain text, html, etc.) with unknown content (known content files may be included as markers) along with a group of Semantic Signatures. The Semantic Signature capture different contents and different attributes of the same content. Semantic Signatures are exposed to each input file to compute the "vector hit"**.** DAT detects semantic features by generating document vectors for each input data file and computing vector hit (within the Semantic Signature classes/clusters) frequencies for each file. DAT generates a *document analysis matrix* as the output, which consists of a *semantic feature vector* for each input file.

Below is the screen shot of the DAT

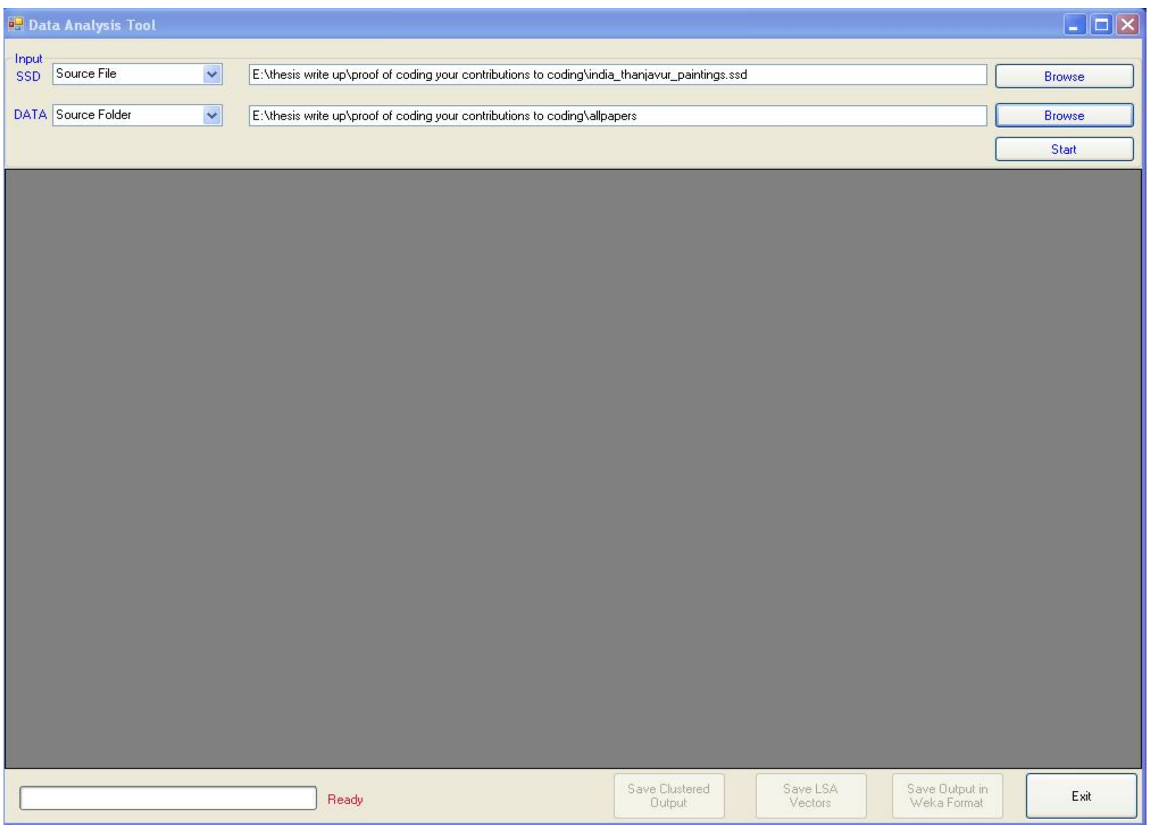


Figure 4.4.1: DAT GUI

There are two input browse buttons with an open file dialog box to choose a file or folder. One Browse click is for the Semantic Signatures. It is analyst's choice as to how many Semantic Signatures are required for the program . Another browse button is for the corpus with unknown and known content files.

Once we define the inputs for DAT, the analyst can just click on the **Start** button. For a fixed Semantic Signature and a fixed file, the program computes the document vectors for the file using the keyword set associated with the Semantic Signature and then computes the number of vectors that fall within the Semantic Signature's cluster. This is done for all Semantic Signature and file pairs.

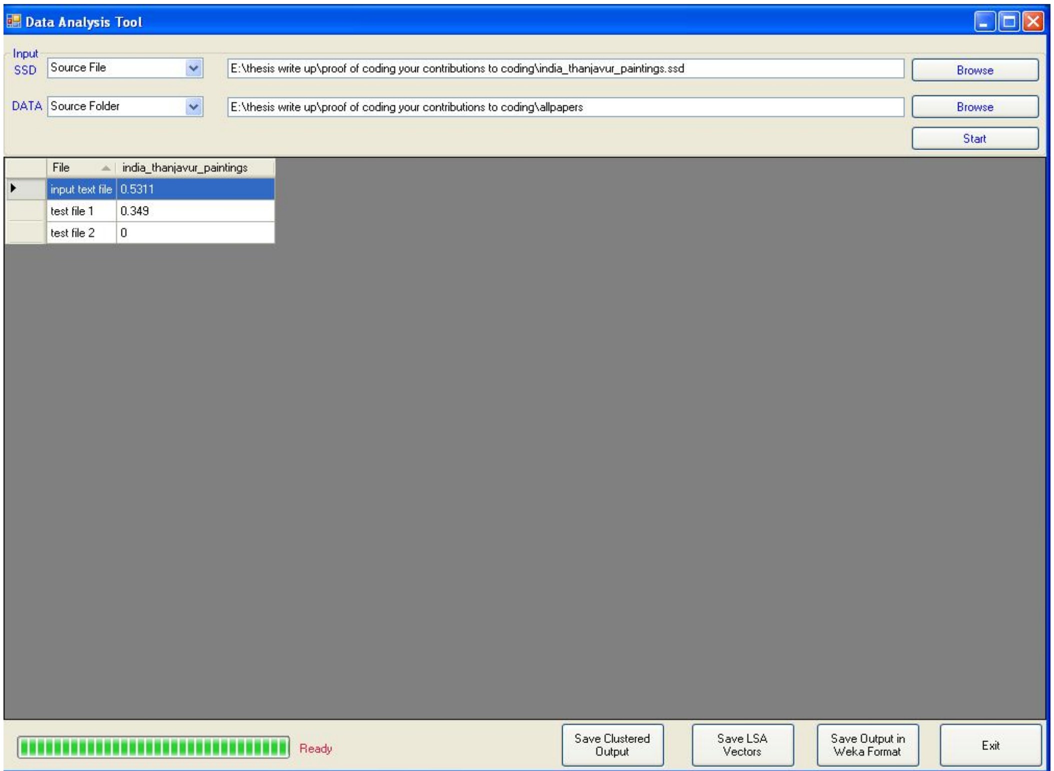


Figure 4.4.2 : DAT showing the Document Analysis Matrix

For the above example, there are two test files, which are the unknown content files and a known content file, included as a marker. Since there is one Semantic Signature, the Document Analysis Matrix, which has Semantic Signatures as columns, has one column. The rows are the input files, making it 3 rows in this example. The Document Analysis Matrix has dimension 3 × 1.

The elements in the matrix indicate the normalized vector hit; the total number of hits is normalized by the length of the file.

If you observe, the numbers as such are relative. For e.g.: The known content file has a hit count of 0.531, relatively high when compared to the test file 1, since the known content file is the source for the Semantic Signature. Also, the test file 1 has a nonzero hit value, as this is a document on Indian Thanjavur paintings and their significance; whereas, test file 2 was a document on an artist who shares his experiences doing such paintings. There wasn't a content which describes India, Thanjavur and paintings together.

4.4.1 DAT Output in .arff Format

The Document Analysis Matrix is saved in the .arff format. This is the format for the input file of WEKA, a data mining open source machine learning software package, used for data clustering.

```
%1 E:\thesis write up\proof of coding your contributions to coding\allpapers\test file 2.txt

%2 E:\thesis write up\proof of coding your contributions to coding\allpapers\input text file.txt

%3 E:\thesis write up\proof of coding your contributions to coding\allpapers\test file 1.txt

@relation 'Data Clustering'

@attribute 'india_thanjavur_paintings' numeric

@data

0

0.5311

0.349
```

Figure 4.4.1.1: Output of DAT in .arff format

This format makes the whole SSMinT package compatible with the WEKA software. Now the Data Analysis Matrix can be further analyzed to cluster or classify the documents.

4.5 WEKA-Data Clustering

WEKA [23] is chosen to be the data clustering tool as it is a open source data mining tool. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. It is also well-suited for developing new machine learning schemes.

WEKA operates on the Document Analysis Matrix. It classifies the corpus of unknown content data files based on semantic content embodied in semantic signatures. Clustering is performed on the semantic feature vectors of the Document Analysis Matrix.

Clustering the Document Analysis Matrix using various clustering algorithms already defined in WEKA can give a better understanding of the corpus of data. The clustering will allow the corpus of data to fall in to well-defined clusters (subsets) on the basis of the Document Analysis Matrix.

5: Stemming Experiment

5.1 Introduction

The definition of stemming was discussed in the preprocessing section of Chapter 2. Stemming is the process of reducing each word to its stem. There are many algorithms which have certain rules in stemming a word to its root. We have chosen the Porter Stemmer algorithm because it's very well known for its simple, unified approach and simplicity.

5.2 Importance of Stemming in Information Retrieval Systems

According to Goldsmith et al. [24], Information Retrieval (IR) is a process involving decision making to identify documents that can satisfy user's need for information. The user's information request is comprised of queries or a search profile plus perhaps some additional information such as weights, etc. The decision making is done by comparing the query term with the index terms (import words or phrases in the document). This decision can be in the form of binary, that is pass/fail or reject/accept, or it can involve a degree of relevance of that document with the query. In most of the cases, structural variations of words have similar semantic interpretations and can be considered as equally relevant when it comes to IR applications. For this reason, numerous stemming algorithms (or stemmers) are employed which attempt to reduce a word to its stem. Stemmers are common elements in web queries analysis and search engines since a user who wants to run a query on "roses", for example, would probably be interested in documents that contain " rose" without the 's' as well. For the purpose of information retrieval, it is not necessary to determine whether the stems generated by the stemming algorithm are valid or not provided that (a) different words with the same 'base meaning' are conflated to the same form, and (b) words with distinct meanings are kept separate.

According to Porter [25], the Porter stemming algorithm (or 'Porter stemmer') is defined as a process for removing the more common morphological and inflexional endings from words in English. Its main use is as part of a term normalization process that is usually done when setting up Information Retrieval systems.

5.3 Objective

The objective of this experiment is to see the effectiveness of stemming in information retrieval as performed by the SSMinT package. Our study starts with stemming as a variable. That is experiments on a large data set are conducted with and without stemming the documents in the data set. We test our tools' performance in document retrieval where the target content is fixed around a topic and the large data set is processed using the SSMinT programs to see if stemming aids in this retrieval.

5.4 Design of the Experiment

We worked with the new version of Reuters corpus (Reuters 2000). According to the press release of Reuters corpus [26], this corpus is made up of 984 Mbytes of newspaper articles in compressed format from issues of Reuters between the 20th Aug., 1996 and 19th Aug., 1997. The number of total news articles is 806,791, which contain 9,822,391 paragraphs, 11,522,874 sentences and about 2 hundred million word occurrences.

The idea is to see how effective stemming is with the SSMinT package. We incorporated stemming as a plug-in to Keyword Tool using the Porter stemming algorithm. Once stemming is chosen in Keyword Tool, stemming is also used in Learner Tool and Data Analysis Tool. The Semantic Signatures can contain the stemmed words as keywords. If the stemming is incorporated, the XML format of SSD and KDF files has an XML node that says stemming has the value "yes", which when read by Learner Tool and Data Analysis Tool, input text files are also stemmed accordingly.

Reuters corpus has topic codes in each of the articles. We chose a topic and randomly picked articles with that topic code. We extracted keyword sets from the randomly picked articles and developed Semantic Signatures with stemming and without stemming. We then collectively exposed these Semantic Signatures to a set of unknown content Reuters articles to see how many articles with the target content were extracted with stemming versus without stemming.

5.5 Methodology/Approach in Choosing Keywords

Each article in the corpus has one or more topic codes attached to it. These topics represent the subject areas of the article. We chose "International Relations" as our topic from which to pick the known content articles. The topic code for "International Relations" is GDIP.

Keyword sets for the stemmed experiment were derived from stemmed data files, and keyword sets for the non-stemmed experiment were drawn from non-stemmed data files. For each article, a pair of keyword sets was chosen carefully: one member of the pair was derived from the non-stemmed data and the other member was derived from the stemmed data. At least one keyword was common to each pair of stemmed and non-stemmed keyword sets; by doing this, we constrained the keyword pairs to target the same semantic content while allowing differences due to stemming/non-stemming.

Also, we chose, for both the stemming and non-stemming case, keywords that did not occur in tight groups, but instead occurred with some space between them in the article. This was necessary because most Reuters articles are short. If the words in a keyword set only appear in a group within a single window, most of the document vectors generated have very few nonzero elements. The resulting set of document vectors does not contain rich enough structure to support analysis.

5.6 Experimental Procedure

- We randomly picked 50 articles with the topic code GDIP from the GDIP topic corpus. We used Keyword Tool to choose keyword sets with and without stemming. From each article, two keyword sets were chosen, one with stemming and another without stemming.
- Semantic Signatures were chosen using Learner Tool with the 50 known content articles and the KDF pairs derived from these articles as input.
- We picked the testing articles from the Reuters corpus at large, keeping in mind not to choose the training articles we used to generate the Semantic Signatures. Twenty folders of the Reuters data were selected to be the pool of testing data. This included 36974 articles.
- The testing articles and the SSD files (50 stemming + 50 non-stemming) were the inputs to Data Analysis Tool (DAT).
- Data Analysis Tool generated the Document Analysis Matrix. The matrix ignores the rows that have zero hits (all zero semantic feature vectors). That is, the articles that did not have any document vector within a Semantic Signature's cluster were ignored, as these articles did contain any of the target content.
- The remaining matrix dimension was 17753 × 100.

5.7 Analysis of the DAT Output

The Document Analysis Matrix is saved by DAT in .arff format so that we can cluster the matrix data in WEKA. Before clustering, to identify the row vector in the matrix, we appended the category code from Reuters of that row's test article. By doing this, these test articles were identified with the topic code, which can aid in identifying the strongest topic in a given cluster. That is, suppose the test article had code GDIP, the code GDIP was appended to that row vector of the matrix. If a test article was a non-GDIP article, then since each article has one or more topic codes assigned to it, the first code that appears in its code list was appended to the row vector of that particular test article in the Document Analysis Matrix.

5.8 Clustering in WEKA

The updated .arff file with the topic codes is fed to WEKA to cluster the matrix data, which is a set of semantic feature vectors. Expected Maximization (EM) was chosen to be the clustering algorithm (in Weka version 3.6), as it is considered to be better than k-means for one main reason – we don't have to guess/provide the number of clusters. EM determines the number of clusters using cross-validation; that is, it internally runs multiple times and picks the number of clusters that resulted in the highest expectation.

Since the Document Analysis Matrix contains 50 stemmed and 50 non-stemmed Semantic Signatures as columns, we wanted to see how the semantic feature vectors clustered with the effect of stemming and non stemming Semantic Signatures individually. The goal being to demonstrate the efficiency of stemmed Semantic Signatures versus non-stemmed Semantic Signatures in retrieving articles with the target content. So EM was run on the semantic feature vectors in two passes: 1) using non-stemmed Semantic Signatures only and 2) using stemmed Semantic Signatures only.

EM	**Clusters**	**Without Stemming**	**With Stemming**
	Cluster 0	**139 (1%)**	**14555(82%)**
	Cluster 1	**81(0%)**	**522(3%)**
	Cluster 2	**249(1%)**	**946(2%)**
	Cluster 3	**582(3%)**	**773(4%)**
	Cluster 4	**704(4%)**	**386(2%)**
	Cluster 5	**897(5%)**	**572(3%)**
	Cluster 6	**14579(82%)**	
	Cluster 7	**339(2%)**	
	Cluster 8	**184(1%)**	

Table 5.1: Expected Maximization clustering on 17753 × 100 Document Analysis Matrix

With the clustering output, we needed to know the contents of the clusters. The technique used was to randomly sample each cluster. We randomly picked 10 articles from each cluster and manually collected specific information from each article. The information gathered is:

- Reuters classifications of the article.
- Manual reading classification – Is the article classified as GDIP by the analyst who designed the Semantic Signatures for capturing "International Relations" content? This is an important distinction, since there is no basis to assume that the Reuters classification matches the classification goals of the analyst.
- Headline of the article.
- Counts of nonzero element values in the semantic feature vector of the article.

For Pass 1:

EM clustering with non- stemmed Semantic Signatures only

Here we considered the 17753 × 100 Document Analysis Matrix, ignoring the stemmed Semantic Signatures and letting the EM clustering algorithm consider only the 50 non-stemmed Semantic Signatures as attributes. There were 17753 row vectors. Weka gave an output of 9 clusters with unequal distribution of articles. The table above describes the cluster distribution. Most of the clusters are relevant to the target content that was captured in the non-stemmed Semantic Signatures, but some clusters stand out as containing a significant number of articles with the target content.

Going forward with the manual sampling, we collected 10 sample articles from each cluster and analyzed manually the collected information for the sampled articles to determine if the cluster is a GDIP cluster (as defined by the analyst). Here is the snapshot of the analysis:

EM – Without Stemming – 9 clusters

			Reuters classification	manual-reading-Is it a GDIP?	Headline of the document	All zero vectors
	Cluster 0	171	C12,C13,CCAT,GCAT,GCRIM	NO	Hong Kong shuts down Internet software pirate	SEVEN 1, ONE 2
		498	GCAT,GCRIM,GDIP,GVIO	YES	Pilot of hijacked Cuban plane returns home	FOUR 1, ONE 2
		1118	GCAT,GDIP,GVIO	YES	Jordan acts against Iraqi diplomats over riots	THREE 1
		2235	M11,MCAT	NO	Israeli shares decline despite rate cut	TWO 1, ONE 3
		3785	C11,CCAT	NO	Kaifa plans US$40 mln venture with IBM	FOUR 1, ONE 2
		5600	C15,C152,CCAT	NO	Sierra Semiconductor jumps on exit plan	THREE 2, ONE 3,6
139		8502	C17,C171,CCAT	YES	DLJ to sponsor APT Satellite's US/HK IPO - sources	SEVEN 1, ONE 2
		12117	C11,C31,CCAT	NO	Bayer plans stronger sales in Czech Republic	ONE 1
40%		13998	E51,E512,ECAT,G15,G158,GCAT,GDIP	YES	EU ministers set to slam U.S. trade laws	THREE 1
		15947	M11,MCAT	NO	Czech CNB-120 index rises 0.7 pts to 868.1	ONE 1
	Cluster 1	39	M14,M143,MCAT	YES	Pakistan issues tender to buy Oct-Dec oil products	FIVE 1, ONE 3
		722	C13,C24,CCAT	NO	Shell asks Turk permission for $2.4 bln LNG plant.	ONE 1
		1443	GCAT,GPOL	YES	Chirac pledges to enforce tough immigration laws	ONE 1, FOUR 2
		2824	GCAT,GPOL,GDIP	YES	Angola expels more than 160 Senegalese	FOUR 1, ONE 2
		3957	C18,C182,C24,C33,CCAT	YES	Hunt to get M.L. Cass stake for acreage	THREE 1
81		5230	E51,E512,ECAT,GCAT,GDIP,M14,M143,MCAT	YES	Gulf traders discount report of Iraqi gas oil sale	ONE 3, FIVE 4
		6992	GCAT,GCRIM,GVIO	YES	E.Berlin spymaster charged with brutal kidnapping	FIVE 1
70%		11330	GCAT,GSPO	NO	TENNIS-GRAF'S FATHER SAYS SHE WAS UNAWARE OF TAX SCHEMES	TWO 1, THREE 2
		14144	GCAT,GDIP,GVIO	YES	Oman says US no-fly coalition has GCC support	TWO 1, FOUR 3
		16208	GCAT,GDEF,GDIP	YES	Bonn commission to look into Libya weapons deal	ONE 1,2
	Cluster 2	353	GCAT,GPOL,GVIO	NO	Burundi army under pressure after rights report.	ONE 2,4 THREE 3
		1103	GCAT,GDIP	YES	Nepal won't help split Tibet, king tells China	EIGHT 1
		2150	GCAT,GVIO	NO	Kashmir hostages reported seen but no confirmation	THREE 1, TWO 2
		4247	C12,C13,C311,CCAT	NO	INTERVIEW - Sahaviriya blasts cheap imports	FIVE 1, ONE 2
249		7062	GCAT,GDIP,GPOL,GVIO	YES	Rights activists accuse Manila of bowing to China	ONE 3,9
		10200	GCAT,GPOL	NO	Polish ruling parties row over minister's sacking	TWO 1
20%		13737	E13,E131,ECAT	NO	Greek core inflation said 8.2 pct yr/yr in Aug	ONE 1
		15666	C13,C31,CCAT,E51,E511,ECAT	NO	India committee to consider foreign media entry	TWO 1, ONE 2
		16652	GCAT,GPOL,GVOTE	NO	Greek PM calls early elections on September 22.	FIVE 1
		17292	GCAT,GPOL	NO	Mandela's ANC defends "just struggle"	FOUR 1, ONE 2
	Cluster 3	87	M14,M143,MCAT	NO	World oil prices fall sharply on profit-taking	ONE 1,4
		110	M14,M143,MCAT	NO	Rampant demand keeps oil price buoyant	ONE 1,4
		430	GCAT,GPOL	NO	Former Russian energy minister to become govt aide	ONE 1
		2301	C15,C152,C21,CCAT	NO	Back to the future at North Sea oil conference	ONE 4
582		6058	C21,CCAT	NO	Japan July vehicle production up 9.5 pct	ONE 1
		7860	GCAT	NO	PRESS DIGEST - Poland - Sept 2	THREE 1
		9946	M11,M13,M132,M14,M142,MCAT	NO	Rangebound bourses cool on second U.S. raid	ONE 2
		11403	M14,M143,MCAT	NO	IPE gas oil ends firm on options expiry, new spec	ONE 2
		13689	C21,CCAT,E31,E311,ECAT	NO	Japan ethylene output hits highest ever Aug level	ONE 1
		16930	C31,C312,CCAT,M14,M141,MCAT	NO	U.S. sorghum weekly export sales highlights -- USDA	ONE 1
	Cluster 4	47	C15,C152,CCAT,M11,MCAT	NO	H-shares up on hopes of Chinese interest rate cut.	FIVE 1
		101	GCAT,GCRIM	NO	China cuts sentence of IMF staffer in graft trial	SIX 1 , ONE 3
704		1690	C13,C33,CCAT	NO	Unify signs licensing pact with Chinese.	THREE 1, ONE 2
		6296	M14,M143,MCAT	NO	Green diesel change leaves IPE gas oil vulnerable	FOUR 2, ONE 1
10%		8625	C13,C24,CCAT	NO	FEATURE - Turkish Islamic banks seek legal changes	ONE 2

Figure 5.8.1: EM clustering Non- Stemming sampling the clusters manually

EM– Without Stemming – 9 Clusters ... Contd.

		9554	M12,MCAT	NO	WORLD BONDS - Data dim U.S. safe-haven demand	ONE 2, ONE 5, THREE 3
		11261	E51,E512,ECAT,GCAT,GDIP	YES	U.S. ex-president Bush says stop anti-China threats	THREE 1, 2 ONE 3
		12699	M14,M143,MCAT	NO	Oil prices extend rally on tighter supply forecast	ONE 6
		13634	M11,MCAT	NO	Shenzhen exchange bids for top China spot	FIVE 1, ONE 2
		16236	C15,C152,CCAT,M11,MCAT	NO	OPINION - INDIA MARKET STRATEGY - BY KOTAK SECURITIES	ONE 6
	Cluster 5	100	E41,ECAT,GCAT,GJOB,GPOL	NO	Netanyahu targets illegal foreign workers	THREE 2
		359	E41,ECAT,GCAT,GJOB,GPOL	NO	Nigeria bans university lecturers' union	ONE 1
		1599	E11,M132	YES	Dollar ends mixed as investors bid up yen	SIX 1
		3977	C33,CCAT	NO	Continental in pact with Business Air	ONE 1
897		6117	GCAT	NO	PRESS DIGEST - Germany - Aug 29	ONE 1,2
		8131	GCAT	NO	PRESS DIGEST - Indian newspapers - Sept 2	FIVE 1
20%		11032	GCAT	NO	RTRS-Australian Broadcasting Corp Afternoon Update	TWO 1
		14199	GCAT,GDIP	YES	Golan settlers say Syria does not want peace	THREE 1,3 ONE 2
		16121	GCAT,GCRIM,GVIO	NO	Israeli agent denies he told of crushing skulls	FIVE 1
		17268	GCAT,GDIP,GVIO	YES	U.N. refugee agency hopes to empty Burundi camp	FOUR 1, ONE 3
	Cluster 6	22	C17,C172,CCAT	NO	Korea Exchange Bank HK mandates HK$500 mln FRCD	ONE 1
		120	C11	NO	INTERVIEW-TREG urges KEPIT holders to accept offer	ONE 1
		412	E12,E13,E131,ECAT	NO	Hungary rate cut tracks inflation fall - analysts	FOUR 1
		1850	GCAT,GPOL,GVIO	NO	Russian, Chechen fighters take time out from war	FOUR 1
14579		4004	C13,CCAT,GCAT,GENV	NO	UK lowers noise limits for three London airports	THREE 1, ONE 3
		6427	C15,C151,CCAT	NO	Lukoil 1st half net profit sharply up	FOUR 1
		10033	C11,C13,C17,CCAT	NO	Britain approves Lloyd's rescue package	FOUR 1
		13009	GCAT,GENV	NO	Al Gore unveils pact to protect forest in Alaska	ALL ZEROS
		14300	M11,MCAT	NO	Blue chips soar; market confident on interest rates	ONE 2
		16800	M11,MCAT	NO	Dutch shares close higher after Buba cuts repo	ONE 1
	Cluster 7	44	GCAT,GDIP	YES	HK democrats see value in contact with Taiwan	THREE 1, ONE 2,3,6,9
		525	M11,MCAT	YES	Tokyo, Hong Kong slide, other Asian markets up	FIVE 1
		1803	GCAT,GDIP	YES	Mandela backs Taiwan, wants links with China too	ONE 3,6,7
339		3797	C24,C31,C312,CCAT	YES	FEATURE - Taiwan steel firms look for greener pasture abroad.	TWO 1, ONE 3,4,8,11
		7530	E51,E512,ECAT,GCAT,GDIP	YES	British minister meets Japan premier after delay	THREE 1 ONE 2,3
90%		9224	C12,C13,CCAT,GCAT,GCRIM,M11,MCAT	NO	China punishes firm, brokerage for irregularities	EIGHT 1, ONE 4
		11211	GCAT,GDIP,GVIO	YES	Russia condemns new U.S. strikes on Iraq	ONE 1
		15355	GCAT	YES	PRESS DIGEST - Taiwan newspapers - September 9	TWO 1,3
		16678	GCAT,GDIP	YES	China summons Ukraine ambassador in row over Taiwan	TWO 1,6 THREE 2 ONE 5
		17750	GCAT,GDIP	YES	China offers to hold political talks with Taiwan	ONE 1,4 TWO 2,9
	Cluster 8	67	E12,ECAT	NO	China says conditions ripe for interest rate cut	ONE 1,6 THREE 5
		402	C11,CCAT	NO	Romania private oil operators plan national co	FOUR 1, TWO 2,6
		1998	C18,C181,CCAT,M11,MCAT	NO	Depressed stock prices spur Singapore takeovers	ONE 1, FOUR 4
		2979	GCAT,GCRIM,GPOL	NO	China dissident says he jumped to avoid beating	THREE 4, ONE 5
184		4917	C11,C13,CCAT,E51,ECAT	NO	India's image problems deter foreign investors	ONE 2, TWO 3, ONE 8
		9899	GCAT,GVIO	NO	U.S. launches cruise missile attacks against Iraq	ONE1,2,6 THREE 3
20%		11254	C17,C171,C18,C181,CCAT	NO	Campbell in $2.5 billion stock buy-back, other initiatives	THREE 5, ONE 6
		13921	GCAT,GCRIM,GPOL	NO	Susan McDougal says Clintons did no wrong	FOUR 4
		15086	C11,C33,CCAT	NO	McDonnell Douglas seeks components from India	TWO 3, ONE 4,8
		16250	E51,E512,ECAT,GCAT,GDIP	YES	India says Pakistan must grant MFN for free trade	TWO 1, ONE 3
		17611	GCAT,GDEF,GDIP	YES	Non-nuclear states to salvage shelved nuke pact	TWO 1, ONE 3

2

Figure 5.8.1.1: EM clustering Non- Stemming sampling the clusters manually…contd

Interpretation: Green highlights are the articles that Reuters classified as GDIP. Red highlights are the articles that Reuters did not classify as GDIP, but they have the target content of "International Relations" as classified by the analyst.

There are a few clusters with richer GDIP content when compared to the other clusters. For example, consider clusters Cluster 1 and Cluster 7, they have 50% or more such articles that talk about international relations. Observe Cluster 6, which is a highly populated cluster with 14579 articles where the sample had no articles with "International Relations" content under the Reuters or the analyst's classification. With this sampling, we learned that some clusters are "pure", and there are clusters with mixed content and for these we cannot for sure state that these clusters represent "International Relations" or not.

We now look at EM clustering with stemming for comparison.

For Pass2:

EM clustering – with stemmed Semantic Signatures only

Here is the snapshot of the EM clustering –with stemmed Semantic Signatures after sampling.

EM - With stemming - 5 clusters

			Reuters classification	manual-reading-Is it a GDIP?	Headline of the document	vectors
	Cluster 0	10	C15,C152,CCAT	NO	First Pac to join Hang Seng London index	all zeros
		110	M14,M143,MCAT	NO	Rampant demand keeps oil price buoyant	ONE 3
		770	GCAT,GCRIM,GVIO	NO	T-Shirts could have financed Trade Center bombing	ONE 1
		3399	C15,C152,C18,C181,CCAT	NO	Essilor rises on option deal	ONE 1
		5007	M14,MCAT	NO	Clean tanker fixtures and enquiries - 1634 GMT	ONE 1
		10331	C15,C152,CCAT	NO	Telebras 1996 net seen over 2.2 bln reais	ONE 1
14555		12243	C31,C312,CCAT, E51,E512,ECAT	NO	Calif exports surge 13.5 pct in first half of 1996	ONE 1
		13277	C31,CCAT	NO	Comet Software in talks to sell rights	ONE 1
		15395	GCAT,GHEA	NO	More than 400 million Chinese suffer low iodine	all zeros
		17307	GCAT	NO	PRESS DIGEST - Bulgaria - Aug 22	ONE 1
	Cluster 1	54	C21,C24,CCAT,GCAT,GDIS	YES	One-fifth N.Korean farmland damaged - Japan academic	TWO 2
		455	C11,CCAT	YES	Gruma, ADM link forges N. American grains powerhouse	FIVE 1
		1008	GCAT,GDIP,GVIO	YES	Hunger strikers vow to maintain Paris protest	ONE 1
		3555	E12,ECAT	YES	BOJ says assessment of Japan economy unchanged	ONE 1, TWO 2
522		7310	GCAT,GCRIM,GPOL,GVIO	YES	Palestinian Authority frees Islamic activists	THREE 1
		11010	C21,CCAT	YES	Alitalia ups Italy-Korea freight capacity	ONE 1
90%		14435	E51,E512,ECAT,GCAT,GDIP,GHEA	YES	Oil deal delay harms Iraqis' health - minister	THREE 1, ONE 2
		16018	E21,E212,E51,ECAT	YES	Mexico, Japan sign credit accords for $960 mln	ONE 1,3,4
		17533	C21,CCAT	NO	Few incidents reported of Hare computer virus	FOUR 1
		17727	C31,C312,CCAT	YES	RTRS-Australia energy exports to Korea seen rising	TWO 1
	Cluster 2	175	C21,E12	NO	FEATURE - Is central planning still right for Singapore	THREE 1
		774	C18,C182,CCAT	NO	Imperial Petroleum to acquire Phonon	ONE 1
946		1487	C11,C12,CCAT	NO	Bratsk Aluminium sees output rising by 1999.	TWO 1 , ONE 2
		2009	C17,C174,CCAT	NO	S&P affirms IBJ, LTCB ratings after tax reassessed.	THREE 1
30%		5060	GCAT,GCRIM,GDEF	NO	French soldiers await trial over sex assault	TWO 1
		8401	E51,E512,ECAT,GCAT,GDIP	YES	Iranian president starts Africa tour	THREE 1
		12162	GCAT,GDIP,GVIO	YES	Pakistan boosts security after U.S. mission stoned	THREE 1, ONE 3
		15069	GCAT	YES	PRESS DIGEST - Bangladesh Newspapers - September 9	THREE 1
		16404	C41,C411,CCAT	NO	Oak Tree Medical names Kedersha CEO	TWO 1
		17300	GCAT,GPOL,GVIO	NO	Yeltsin sees Chechnya as bleeding wound for Russia	TWO 1
	Cluster 3	47	C15,C152,CCAT,M11,MCAT	NO	H-shares up on hopes of Chinese interest rate cut	ONE 1
773		179	C18,C181,CCAT	NO	Unilever looks to buy Israel's Sunfrost -paper	ONE 1, ONE 2
		724	C33,CCAT	NO	GE, Tandem in pact for electronic data system in China	ONE 1
50%		1187	M14,M141,MCAT	NO	Pakistan vegetable oil prices seen holding firm	ONE 1, ONE 4
		3333	M11,MCAT	YES	Africa Israel boosts slow Israeli share market	ONE 2
		7427	M13,M131,MCAT	NO	INDIA - Primary dealers quotes on securities Aug 31	TWO 1
		9973	M14,M143,MCAT	YES	Global oil prices retrace some Iraq-inspired gains	ONE 2, ONE 4
		10914	C21,C24,CCAT	YES	Syria expects to double oil, gas reserves	TWO 2, ONE 3
		12772	GCAT,GDIP	YES	Slovenia to bid for UN Security Council seat	ONE 1
		16342	GCAT,GDIP	YES	Syria worries Israeli intelligence - legislator	TWO 1, ONE 2,5

3

Figure 5.8.2 : EM clustering with stemming – manual sampling

Figure 5.8.2.1: EM clustering with stemming – manual sampling…contd

Interpretation: Here in EM clustering considering only stemmed Semantic Signatures, there are fewer ambiguous clusters when compared to EM clustering without stemming.

In these 5 clusters, a few clusters are rich in "International Relations" content, for example, consider clusters 1 and 4 and there is one cluster where the sample had no articles with "International Relations" content under the Reuters or the analyst's classification. Yet, here also, we couldn't determine for some mixed clusters whether they represent "International Relations" or not.

Learning from this experiment, we decided to prune Semantic Signatures that were not involved in defining an "International Relations" cluster. By doing so, we reduced the dimensionality of the clustering. These unnecessary Semantic Signatures cause extra dimensions in the mathematical space and they tend to give rise to ambiguous clusters. This motivated us to perform a second iteration by using only specific Semantic Signatures that contributed to the "International Relations" rich clusters.

5.8.2 Dimensionality Reduction- Iteration 2

We considered only those clusters that have 50% or more "International Relations" content articles in the sample. We examined the semantic feature vectors of the articles sampled from these clusters and retained only the Semantic Signatures that had at least one hit by a document vector from an article in the cluster. The number of Semantic Signatures was brought down to 33 from 100 (18 non-stemmed + 15 stemmed). DAT processed the corpus on these 33 Semantic Signature and gave a new Document Analysis Matrix with dimensions 16834 × 33. 16834 articles were retrieved by these 33 Semantic Signatures, which was a lot less than the original

17753 from Iteration 1. We then analyzed the clusters individually to see if there is a variation in the distribution.

EM clustering without stemming – pruning the Semantic Signatures – Iteration 2

Here we have only 18 non-stemmed Semantic Signatures to consider and EM clustering algorithm in Weka was run on the 16,834 files to generate 8 clusters. After manual sampling the clusters, we had better, purer clusters when compared to Iteration 1. Here is the snapshot of the analysis:

Pass 2 – Pruning SSD's – EM – Without Stemming – 8 clusters

			Reuters classification	manual-reading-Is it a GDIP?	Headline of the document	All zero vectors
	Cluster 0	67	GCAT,GCRIM	NO	FEATURE - Philippines steps up fight against paedophiles	All zeros
		1010	M14,M141,MCAT	NO	LCE cocoa slips below support as longs sell	one 2
		3192	M11,MCAT	NO	SOLIDERE shares mixed in Beirut	All zeros
5147		5168	C15,C151,CCAT	NO	Italy's Monte Paschi bank sees jump in H1 results	one 2
		7071	C31,CCAT	NO	UK's Sunday shoppers favour food superstores	ONE 1
		9762	GCAT,GDIP,GPOL	YES	Sandinistas may change anthem to smooth image	ALL ZEROS
20%		11117	C12,C41,C411,CCAT,GCAT,GCRIM	NO	Sanwa Bank image tainted by embezzlement charge	ONE 3
		12882	E12,ECAT	NO	Infometrics wants output stability in RBNZ aims	ONE 1
		14766	C21,CCAT,E31,ECAT	YES/NO	China provinces focus on same industries	ONE 2
		15786	GCAT	NO	PRESS DIGEST - Sweden - Aug 21	ALL ZEROS
	Cluster 1	120	C18,C181,CCAT	NO	GWR says to merge with Classic FM	THREE 1
		465	E11,ECAT	NO	Chile's economy grows 7.9 pct in first half 1996	FOUR 1
		1502	C13,CCAT,E21,E211,E51,E512,ECAT,	NO	Hanoi to get tough with import-export tax dodgers	FOUR 1
		3403	E71,ECAT	NO	Taiwan July index of leading indicators down	FOUR 1
6228		6866	M12,M13,M131,MCAT	NO	Canadian bonds weaker at early close on U.S. data	FOUR 1
		9104	E21,E212,ECAT,M12,MCAT	NO	Egypt offers 4 bln pounds worth of treasury bonds	FOUR 1
		11771	C15,C152,CCAT	NO	Vickers eases on worries about tank orders	FOUR 1
		13647	NULL	NO	Richemont rises, MIH slips on Nethold deal	THREE 1, ONE 2
		14899	C17,C172,CCAT	NO	FINNISH CO-OPERATIVE BANKS LAUNCH FIM 100 MLN BOND	FOUR 1
		16312	C33,CCAT	NO	Airport Systems gets $1.2 mln in contracts	FOUR 1
	Cluster 2	417	GCAT,GPOL	NO	Ukraine appoints new investment agency head	ONE 1
		981	GCAT,GPOL	NO	Gaza journalists briefly boycott cabinet meeting	ONE 1
		1538	C41,C411,CCAT	NO	CBT Group names Buckley president, COO	TWO 1
		2800	E51,E512,ECAT,GCAT,GDIP	YES	Japan's Hashimoto meets Peru's President Fujimori	ONE 1
669		5997	GCAT,GCRIM,GPOL,GVIO	NO	Apartheid generals offer to help truth commission	TWO 1
		9768	GCAT,GVIO	NO	Rights group urges safe return of Colombian troops	ONE 2
20%		11957	C31,C311,CCAT,E51,E512,ECAT	NO	Japan buys $4.87 bln European auto parts in 95/96	ONE 1
		13446	GCAT, GVIO	YES	Kuwaitis welcome reported bid to oust Saddam	TWO 1
		15566	C41,C411,CCAT	NO	Radnet taps Lotus exec for CEO post	TWO 1
		16291	C18,C181,CCAT	NO	Seagram to merge its two U.S. wines cos	TWO 1
	Cluster 3	112	C17,CCAT	NO	Mercury One 2 One seeking to extend debt	FOUR 2
		1854	GCAT,GPOL,GCRIM,GVIO	YES	Cambodia's Ieng Sary says he's no mass murderer	THREE 3, ONE 4
		3928	M11,MCAT	NO	Bank selloff, energy news move Hungary's OTC market	FOUR 2
		6290	C12,CCAT,GCAT,GCRIM, GDIS	YES	French magistrate opens probe into TWA crash	FOUR 2
		8001	C11,CCAT	YES	Samsung plans Australian investment spree	FOUR 2
2281		8732	C21,CCAT	NO	Japan microchip makers shift to 64-megabit DRAMs	THREE 2, ONE 3
		10088	M12,MCAT	NO	Dutch closing debt market report	FOUR 2
40%		11558	C11,C18,C181,CCAT	YES	India's HPCL said seeking Exxon stake in new unit	ONE 1,3 THREE 2
		12655	C22,CCAT	NO	Daiwa expands US Treasuries' trade into cyberspace	FOUR 2
		13462	GCAT,GVIO,GWEA	NO	Hurricane cleanup under way in North Carolina	THREE 2, ONE 3

5

Pass 2 – Pruning SSD'S – EM – Without Stemming – 8 clusters ... Contd.

	Cluster 4	150	M141	NO	China is in no hurry to import palm oil	THREE 1,3 ONE 4, 7
		343	GCAT,GPOL,GVIO	NO	Burundi army under pressure after rights report	ONE 2,4 THREE 3
		1850	C11,C15,C152,CCAT	YES	INTERVIEW - Asia energy investment still wary.	TWO 4,5
		2174	GCAT,GPOL,GVIO	YES	Africans set up camp, trade tales of French raid	ONE 2,4 THREE 3
		6681	M14,M143,MCAT	YES	Asia jet under pressure, but gas oil could cushion	ONE 1,5 THREE 2
		8040	C17,C174,CCAT	YES	RTRS-S&P upgrades Pioneer Intl to A-minus	ONE 6
269		10877	GCAT,GPOL,GVIO	YES	S.Africa's Buthelezi slams ANC, claims innocence	FIVE 1 ONE 3
70%		13775	C15,C152,CCAT	NO	Broderbund warns of lower profits	THREE 4 ONE 6
		15036	C31,CCAT,M14,MCAT	YES	Ex-UK air cargo stable, market hesistates over rate rises	THREE 5 ONE 6
		16539	C11,C24,CCAT	YES	Taiwan CPC to meet partners over Ecuador dispute	ONE 2,3 THREE 4
	Cluster 5	57	C31,C311,C312,C33,CCAT	YES	Venalum plans to ship aluminium to Japan - sources	ONE 3
		900	GCAT,GDIP	YES	Taiwan urges China talks, vows to press diplomacy	THREE 1, TWO 2, ONE 4,5
		2463	E51,E512,ECAT	YES	Taiwan-China trade edges up in Jan-June yr/yr	THREE 4, ONE 6
		5677	GCAT	YES	PRESS DIGEST - Taiwan newspapers - August 29	THREE 1, TWO 2
559		8553	C18,C183,CCAT	NO	Russia to sell Svyazinvest, Transneft stakes	TWO 1, THREE 2
		10881	GCAT,GCRIM,GPOL	NO	Madagascar court confirms president's impeachment	FOUR 1, TWO 2
		12248	C31,CCAT	YES	Northwest sees drop in fish shipments	ONE 1, FOUR 2
80%		14644	GCAT,GPOL	YES	FEATURE - China lauds Mao 20 years after death	FOUR 1, THREE 2, ONE 3
		15710	C15,C152,CCAT	YES	COSCO down on Taiwan/China shipping link	SIX 2, TWO 4
		16829	GCAT,GDIP	YES	Media say Taiwan, Ukraine agreed office exchange	ONE 1,4,8 TWO 6
	Cluster 6	1096	GCAT,GDIP	YES	Cuban exile flights to spot rafters resume	FOUR 1
		2277	E51,ECAT,GCAT,GDIP	YES	Indonesia, Argentina agree to increase trade	SIX 1
		3777	E51,E512,ECAT,GCAT,GDIP	YES	Clinton to keep up trade pressure on Japan - Tyson	SIX 1
		4332	C41,C411,CCAT	YES	AW Computer names McMullin as chairman	FIVE 1
		8091	GCAT,GDIP	YES	Singapore's Lee Kuan Yew in China, to meet Jiang	FOUR 1 ONE 2
135		10013	GCAT,GDIP,GVIO	YES	Clinton not getting world backing on Iraq	SIX 1
		10852	M13,M132,MCAT	YES	Comatose dollar gets fillip from Yeltsin news	FIVE 1, ONE 2
100%		12110	GCAT,GPOL,GPRO	YES	Russians skilled at bypass, Yeltsin may need more	SIX 1
		13089	GCAT,GDEF,GDIP	YES	Kohl sees "active" Yeltsin resolving NATO issue	FOUR 1, ONE 2
		14807	C16,E12,M132	YES	Dollar falls against mark, gains on yen	SIX 1
	Cluster 7	109	M14,M143,MCAT	YES	Rampant demand keeps oil price buoyant	ONE 1,4
		1001	C151,E21,E211	NO	India Sanghi Poly 95/96 net falls 67 pct	ONE 1
		2002	E51,E512,ECAT,GCAT,GDIP	YES	Iraqi oil delegation coming to UN in New York	ONE 1
		3502	C181	NO	MFS-WorldCom deal opens gate for resellers	TWO 1
1546		6077	E51,E513,ECAT	NO	Polish currency reserves rise to $17.8 bln in July	TWO 1
		7412	GCAT,GDIP	YES	Iranian president starts Africa tour	ONE 3
50%		10152	M12,M13,M131,MCAT	NO	INDIA-Primary dealers quotes on securities-Sept 4	ONE 1
		11118	GCAT	YES	PRESS DIGEST - Pakistan - September 5	ONE 1
		13742	GCAT,GSPO	YES	CRICKET-EX-ENGLAND CRICKETER BOTHAM TO APPEAL IN LIB	ONE 1
		15742	C15,C152,CCAT	NO	High costs, inefficiency hit China light industry	TWO 1
		16771	M14,M141,MCAT	YES	Pakistan has Sept option on 75,000 T Indian sugar	ONE 1

6

Fig 5.8.3: EM clustering without stemming – after pruning – iteration 2

Interpretation: Though the number of clusters remains almost the same, the clusters have more purity. Clusters 1, 4, 5, 6, 7 are examples of pure clusters. The number of mixed clusters is greatly reduced.

EM clustering with stemming – pruning the Semantic Signatures - Iteration2

Here we ran EM clustering on 16834 files considering 15 stemmed Semantic Signatures that contributed towards "International Relations" clusters. Here is the snapshot of the manual sampling:

Pass 2 – Pruning SSD's – EM – With stemming – 3 clusters

			Reuters classification	manual-reading-Is it a GDIP?	Headline of the document	All zero vectors
	Cluster 0	39	M14,M143,MCAT	YES	Pakistan issues tender to buy Oct-Dec oil products	ONE 1,4
		370	GCAT,GDEF,GDIP	YES	Ukraine denies Taiwan pilots tested jet fighter	TWO 1
		1301	GCAT,GPOL,GVIO	YES	Manila to sign Moslem peace accord Sept 2	ONE 1
		3358	C24,CCAT,GCAT,GDIS,GENV	YES	N.Korea says over 270,000 ha of farm land submerged	TWO 1
905		7274	GCAT,GPOL	YES	Socialist Jospin targets French coalition's record	TWO 1
		8073	GCAT,GPOL,GDIP	YES	Rifkind says optimistic about HK transfer to China	ONE 1,2
90%		10597	C31,C311,CCAT,E11,E51,E512,ECAT,GCAT,GDIP,GVIO	YES	Japan economy not seen suffering much from US-Iraq row	TWO 1, ONE 4
		12732	GCAT,GVIO	YES	Allied planes report limited Iraq radar activity	THREE 1, ONE 3
		14070	M14,M141,M142,M143,MCAT	YES	RTRS-Australian Commodities Roundup - Sept 10	THREE 1
		15714	C31,CCAT	NO	In dull car market, rental firms boom in Shanghai	TWO 1, ONE 3
	Cluster 1	226	GCAT	YES/NO	PRESS DIGEST - RUSSIA - AUG 23	ONE 1
		3997	GCAT	YES/NO	RTRS-PRESS DIGEST-Australian General News -Aug 29	ONE 1
887		5521	GCAT,GENV,GVIO	YES/NO	German anti-nuclear activists in pantomime protest	ONE 2
		9403	GCAT	YES/NO	PRESS DIGEST - Israel - Sept 3	ONE 1,4
		8748	M14,M141,MCAT	YES	China seen watching soymeal as prices rally	TWO 1 ONE 2,3
5%		10362	GCAT,GDIP,GVIO	YES	U.S. missiles hit Iraqi targets again	TWO 1 ONE 2
		13087	GCAT,GDIP,GPOL	YES	Yeltsin and Kohl meet amid Russian power debate	ONE 1
		14722	C24,CCAT	YES/NO	Conn., Northeast Utilities face off over power costs	ONE 1,5
		15021	E51,E512,ECAT,GCAT,GDIP	YES	U.S. to clarify Iran-Libya sanctions in September	THREE 1
		16767	C13,C22,CCAT	YES/NO	India approves Birla Comm phone deal - news agency	TWO 1
	Cluster 2	20	C11,C21,CCAT	NO	Garuda Indonesia to increase Batam direct flights	ALL ZEROS
		605	GCAT,GCRIM	NO	Belgian police dig into night in child sex case	ALL ZEROS
		1250	GCAT,GDIP	YES	Southern Africa puts faith in Mandela's hands	ALL ZEROS
		2612	C11,C31,C312,CCAT	NO	Sun aims to boost Latam business	TWO 2
15042		4033	C13,C31,C311,CCAT	NO	INTERVIEW - Sahaviriya blasts cheap imports	TWO 1
		7706	C12,C33,CCAT,GCAT,GCRIM	NO	Four released in Frankfurt airport probe	ALL ZEROS
20%		10369	M14,M141,MCAT	YES/NO	U.S. exporters sell 100,000 T wheat to Sri Lanka	ONE 1
		11341	C13,C21,CCAT	NO	Aflatoxin found in Texas grain sorghum	ONE 2
		12952	E11,E12,E21,E211,ECAT	NO	Japan Kubo says will work for stronger recovery	TWO 1
		15354	C17,C172,CCAT	NO	Freddie Mac sets $300 mln of REMICs	ONE 1
		16017	M11,MCAT	NO	UAE stocks unmoved, but selling pressure seen	ALL ZEROS

Figure 5.8.4 EM clustering with stemming – after pruning – iteration 2

Interpretation: The number of clusters is 3. Observe, the number of ambiguous clusters drastically reduced, giving rise to more pure clusters with essentially no ambiguity. Also, the largest cluster (cluster 2) has very less occurrences of "International Relations" content articles and very pure clusters 0 and 1.

5.9 Final Conclusions

The stemming experiment included two goals (a) to prove the effectiveness of stemming when used with the SSMinT software tools, (b) to prove the effectiveness of Semantic Signature pruning and dimensionality reduction in the data analysis preformed on the output of the SSMinT software.

In the second iteration, effectiveness of stemming was evident in rendering fine unambiguous clusters. Thus, for applications like information retrieval, stemming proved effective as it can retrieve pure clusters of documents with the target content.

Dimensionality reduction is another effective technique, which aided in proving the effectiveness of stemming as with the unnecessary dimensions in the vector space we couldn't analyze the experiment. The unnecessary dimensions introduced noise into the data.

Stemming gave fewer mixed clusters when compared to non-stemming in both the Iteration 1 and Iteration 2 experiments. Also, from Iteration 1 to Iteration 2, the number of clusters was significantly reduced and the clusters in Iteration 2 were more pure.

Semantic Signature pruning and dimensionality reduction proved to be a powerful tool that is worthy of further investigation in the context of our SSMinT software package.

6: Semantic Sensitivity Experiments

6.1 Introduction

English, in both its written and oral forms, is a difficult language to learn. One of the main reasons for this difficulty is that many of the words have more than one meaning and these meanings vary according to the context in which they are used.

Since English is finite and one of the main limits is the number of words it contains, it has become necessary that a single word take on more than one meaning. This helps to convey the many nuances of human experiences. Thus, the meaning of a word could change based on the context in which it is used.

As mentioned in [27] by Svedman, the term "Semantic Sensitivity" was first coined by Sidney Rauch (1967) in his article on teaching disadvantaged children. According to him, "semantic sensitivity" refers to awareness that words have more than one meaning and the particular meaning implied varies with the context.

6.2 The Study

In this study we have proposed experiments to see how SSMinT responds to the semantic sensitivity nature of the English language. We can say that the semantic sensitivity nature of certain words vary according to the context they occur; therefore, the order in which the keywords are placed and their proximity to each other implies a certain orientation of document vectors in multi-dimensional space.

Our study starts with a set of experiments that process documents that contain closely related topics (throat singing and throat cancer), which are linked to one another and yet are different in their usage and genre. Such documents were carefully chosen subjected to the SSMinT data mining tool. Different experiments were conducted to prove that SSMinT can identify the subtle differences between these two datasets.

Another interesting set of experiments are performed with the 10-K filings of publicly held companies found in U.S Securities and Exchange Commission (www.sec.gov). Here we chose retail market companies that went bankrupt in 2009 and extract their annual filings. On the other side, we chose comparable retail market companies that did not go bankrupt in 2009. The formats of all 10-K reports are similar in a boilerplate fashion; the goal was to assess the utility of SSMinT in identifying companies that will go bankrupt from the text content of 10-k reports.

6A: Throat Singing/Throat Cancer Hierarchical Classification

6A.1 Design/Set up

The Semantic Sensitivity Analysis Experiment was designed to validate if SSMinT can identify the minute differences between closely related documents. To test the scale of sensitivity, we initially analyzed a small pool of data. Throat Singing and Throat Cancer are our first chosen examples of context. These are certainly closely related topics as they concern the stress on the throat and the symptoms caused by either throat singing or by diseases like throat cancer.

6A.2 Approach

- To initialize the experiment, we collected 4 papers from each genre. Throat Singing papers include the topics: a study on throat singing, a study of a specific type of singing, singers from Tuva, blending vocal music, overview of types of throat singing. Throat Cancer papers include the topics: medical and non-medical papers that concern throat cancer in various aspects such as definitions, causes, risks, treatments, and demographics. From each of these 8 papers, keyword sets were chosen that could significantly extract the content of the papers. The most common words like "throat", "cancer", "singing" were ignored in order to assess the sensitivity of the SSMinT methods in differentiating between closely related topics on the basis of semantic structure and keyword sets designed to capture the content specific to each paper (for example a keyword set includes "tumor", "surgery" and "treatment"). The ignored words could easily differentiate between Throat Singing and Throat Cancer papers via a simple keyword frequency count. The keyword sets were carefully built to capture certain content from the learning papers they came from. Then each of these keyword sets was exposed to their own root paper to generate document vectors and develop the Semantic Signatures in Learner Tool that capture specific content within the root papers.

- The Semantic Signatures have the power of identifying the target content in any text, When compared to the bag-of-words approach, bag-of-words can merely associate the frequencies of the keywords, but cannot recognize the structure of keywords as they appear in the text. For the 8 papers (4 from Throat Singing and 4 from Throat Cancer), 3 keyword sets per paper were generated. Correspondingly, 3 Semantic Signatures were generated for each paper. Twenty-four Semantic Signatures, each designed to capture different content, were developed and made ready for further experiments.

6A.3 Experimental Procedure

6A.3.1 Experiment 1:

We used Ke yword Tool and Learner Tool to deve lop the 2 4 Sem antic Signatures. In an initia l experiment, these 24 Sem antic Signatures were exposed to their 8 root papers in the third tool – Data Analysis Tool (DAT). DAT generates a matrix called the Docum ent Analysis Matrix with the hit coun ts for the Sem antic Signatures (nu mber of hits by the docum ent vectors of the 8 papers).

	SSD1																							SSD24
	0	0	0	0	0	0	1	0	0	0	11	0	20	12	19	1	0	11	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	3	3	0	1	1	11	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	1	0	14	9	0	0	0	29	9	8	2	0	0	0
	0	0	0	0	0	0	0	0	0	0	6	0	12	16	0	1	0	6	0	2	0	2	1	3
	11	29	8	0	0	1	3	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	9	9	11	0	0	1	0	5	4	0	0	0	1	0	0	0	0	0	0	0	0
	0	7	0	0	0	0	19	21	32	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	1	0	4	0	1	0	0	2	0	10	4	7	0	0	0	0	0	0	0	0	0	0	0

	Types
Throat Singing	Overtone Singing
	Study on throat Singing
	The throat singers of tuva
Throat Cancer	1985 FINDINGS ON HEALTH PROMOTION AND DISEASE PREVENTION
	Ear, Nose, and Throat Cancer : Ultrasound Diagnosis of Metastasis to Cervical Lymph Nodes
	CANCER COVERAGE AND TOBACCO ADVERTISING IN AFRICAN-AMERICAN WOMEN'S POPULAR MAGAZINES
	Cancer - throat or larynx -Overview

Fig.6A.3.1.1: The Document analysis matrix generated from the 8 papers with 24 Semantic Signatures

This Docum ent Analysis Ma trix is an 8 × 2 4 matrix with sem antic f eature vec tors as rows. Observe the Semantic Signature that is pointed to by the red arrow. This S emantic Signature, for example, has hits generated by Throat Singing and Throat Cancer papers, though, it was derived from a Throat Cancer genre root paper. The pa per from which it was derived was about a survey in the African-American women population that had lot of non-anatomy terms and was about the throat in ge neral. Thus, ther e a re s ome hits by docum ent vectors from Throat Singing papers also.

6A.3.2 Clustering

To analyze the Document Analysis Matrix with clustering, we have used WEKA, a popular open source machine learning software available at (www.cs.waikato.ac.nz/ml/weka/). Weka has several types of clustering techniques that we can use to analyze the output of DAT.

6A.3.2.1 Result of Simple K-means Using Euclidean Distance and Two Clusters

- *Cluster 1* has 4 papers. 3 papers are from Throat Singing and one paper from Throat Cancer. The paper from Throat Cancer is a non-scientific paper that is a survey on a certain sect of people.

- *Cluster 2* has 4 papers. 3 papers are from Throat Cancer and 1 paper is from Throat Singing. The paper from Throat Singing is "Overtone singing".

From the above basic clustering, we see that the clusters reflect the core genres. Though, each of genres had one paper in exchange, it seemed interesting to investigate the distribution of the clusters.

The Throat Cancer paper that was clustered with Throat Singing as it was a non-anatomy paper that was discussing the "cancer coverage and tobacco advertising in African-American women's popular magazines". Similarly, one Throat Singing paper was grouped with the Throat Cancer papers as this discusses the technicalities with the overtone singing techniques which were mostly about singing with the throat under stress and also about the consequences.

6A.3.2.2 Simple K-means Using Cosine Distance and Two Clusters

- *Cluster 1* has 5 papers. This cluster includes all 4 papers from the Throat Cancer genre and 1 paper from Throat Singing genre, which is about types of throat singing.
- *Cluster 2* as 3 papers. This is a pure cluster from Throat Singing.

6A.3.2.3 Cobweb Clustering with Eight Papers

Cobweb clustering shows the hierarchical breakdown of the papers and the sublevels are categorized with the similar cluster orientation.

Leaf 1 has one paper standing apart from the rest of the papers and is a Throat Cancer paper, but is a non-technical paper, mostly about a survey on the African-American women with throat cancer.

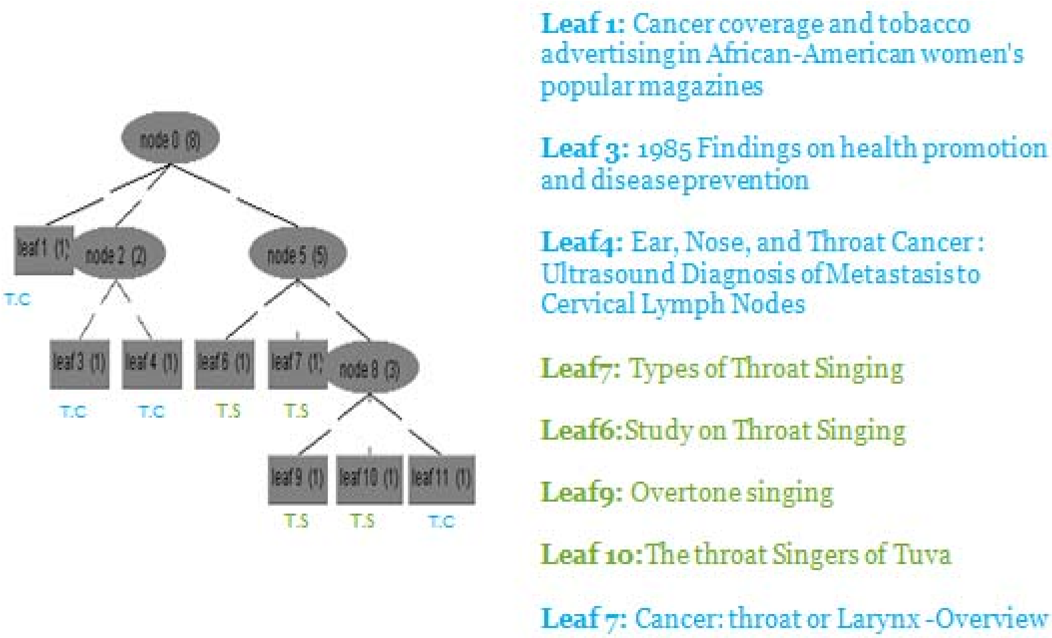


Fig6A.3.2.3: Cobweb clustering on the 8 papers

Node 2 has two Throat Cancer papers, Leaf 3 and Leaf 4. These two leaves are grouped together in one Node (Node 2) and at this level Leaf 1 also is a Throat Cancer paper. Leaves 3 and 4 have the medical papers in Throat Cancer which have similar cluster orientations.

Node 5 mostly contains Throat Singing papers. The classification is spread into leaves and sub nodes. Leaf 6 and Leaf 7 are generic introduction and study on throat singing and types of throat singing. Node 5 splits into a sub node Node 8, which has more approaches and specific definitions about throat singing. It is interesting to note that Leaf 11 includes a Throat Cancer paper which gives an overview on cancer and its Semantic Signature orientation matches that of the remaining Throat Singing papers which discuss the technique and the stress on the throat and its effects.

6A.3.3 Experiment 2

To enhance the experiment and to test the sensitivity of SSMinT, in addition to the learning set of Experiment 1, we included 12 papers, 6 from Throat Cancer and 6 from Throat Singing in the corpus. We used the 24 Semantic Signatures sets as in Experiment 1. Thus, to develop a new matrix we ran DAT on these 20 papers (8 root papers + 12 additional papers) with our 24 Semantic Signatures. The output matrix is a 20 × 24 matrix with numbers indicating hits of the corresponding Semantic Signatures for each paper.

	SSD1	SSD2	SSD3	SSD4	SSD5	SSD6	SSD7	SSD8	SSD9	SSD10	SSD11	SSD12	SSD13	SSD14	SSD15	SSD16	SSD17	SSD18	SSD19	SSD20	SSD21	SSD22	SSD23	SSD24
	0	1	0	8	0	1	1	0	1	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	2	0	0	0	0	0	1	2	0	7	3	4	2	0	0	0	0	0	0	0	0	0	0
	0	0	0	1	0	8	0	0	0	0	6	6	0	0	0	0	0	0	0	0	0	0	0	0
	0	12	0	0	0	0	5	23	14	0	29	0	4	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	1	0	8	0	0	0	0	14	6	1	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Main Paper	11	29	8	0	0	1	3	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
Main Paper	0	0	0	9	10	13	0	0	1	0	5	5	0	0	0	1	0	0	0	0	0	0	0	0
Main Paper	0	7	0	0	0	0	19	21	32	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0
Main Paper	0	1	0	4	0	1	0	0	2	0	10	4	7	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	15	11	2	0	0	14	0	5	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	16	0	24	10	0	0	0	2	2	10	0	0	0	0
	0	0	0	1	0	0	0	0	0	0	4	0	18	8	2	1	0	40	1	4	0	0	0	0
Main Paper	0	0	0	0	0	0	0	0	0	0	0	0	3	3	0	1	1	11	0	0	0	0	0	0
Main Paper	0	0	0	0	0	0	0	0	0	0	1	0	14	9	0	0	0	29	9	8	2	0	0	0
Main Paper	0	0	0	0	0	0	0	0	0	0	6	0	12	16	0	1	0	6	0	2	0	2	1	3
	0	0	0	0	0	0	4	0	0	0	13	0	22	2	6	2	0	15	0	2	0	0	0	1
Main Paper	0	0	0	0	0	0	1	0	0	0	11	0	20	12	19	1	0	11	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	19	0	22	2	2	0	0	20	0	1	0	0	0	0

Throat Cancer

Throat Singing

Fig 6A.3.3: The Document Analysis Matrix generated from the 20 papers with 24 Semantic Signatures

6A.3.4 Clustering

6A.3.4.1 Simple K-means with Three Clusters

Cluster 1:

PAPER TITLES:

- Study on throat singing

Cluster 2:

PAPER TITLES:

- Cancer cov erage and tobacco ad vertising in African-Am erican wom en's popular magazines
- Diet in the etiology of or al and pharyngeal cancer am ong wom en from the southern United States
- New throat cancer treatment
- Perceived risks of certain types of cance r and heart disease am ong Asian Am erican smokers and non-smokers
- Smoking and cancer of the mouth, pharynx and larynx
- Harmonic overtone singing
- Ear, nose, and throat cancer: ultrasound diagnosis of metastasis to cervical lymph nodes

- Drinking levels, knowledge, and associated characteristics, 1985 NHIS findings
- Quality of life 5-10 years after primary surgery
- Cancer - throat or larynx
- Oral mucositis in cancer therapy

Cluster 3:

PAPER TITLES:

- A study of the blending of vocal music with the sound field by different singing styles
- Inuit thro at-games and Siberian throat singing: a com parative, historical, and semiological approach
- Mongolian conceptualizations of overtone singing
- Overtone singing
- Study on throat singing
- The throat singers of Tuva
- Tuvan throat singing
- Types of throat singing
- What is throat singing

Three clusters were chosen to see the classification of the papers with Semantic Signatures in the multi-dimensional space. By selecting m ore than 2 clusters, we are giv ing scope to the clu sters that m ay not be pure and have docum ents with similar orientation. Thus , sensitivity can be thoroughly explained with the distri bution of the papers in to the clusters that have sim ilar orientation.

Cluster 1: "Study on throat singing" stood distin ct without any grouping. This paper is about certain methodologies of throat singing.

Cluster 2: A ll the Throat Cancer papers were grouped together; "Harm onic overtone singing", which is about Throat Singing is also grouped with these papers.

Cluster 3: All remaining Throat Singing papers are grouped together forming a pure cluster.

6A.3.4.2 Simple K-means with Four Clusters

Cluster 1:

PAPER TITLES:

- What is throat singing
- Types of throat singing
- Tuvan throat singing
- A study of the blending of vocal music with the sound field by different singing styles
- Inuit throat-games and Siberian throat singing: a comparative, historical, and semiological approach
- The throat singers of Tuva
- Mongolian conceptualizations of overtone singing

Cluster 2:

PAPER TITLES:

- Perceived risks of certain types of cancer and heart disease among Asian American smokers and non-smokers
- Diet in the etiology of oral and pharyngeal cancer among women from the southern United States
- Cancer coverage and tobacco advertising in African-American women's popular magazines

Cluster 3:

PAPER TITLES:

- Study on throat singing

Cluster 4:

PAPER TITLES:

- Smoking and cancer of the mouth, pharynx and larynx
- Oral mucositis in cancer therapy
- Ear, nose, and throat cancer : ultrasound diagnosis of metastasis to cervical lymph nodes
- Quality of life 5-10 years after primary surgery
- Drinking levels, knowledge, and associated characteristics, 1985 NHIS findings
- New throat cancer treatment
- Harmonic overtone singing
- Overtone singing

- Cancer - throat or larynx

The groupings show sensitivity to the subdivisions of content.

Cluster 1: A purely Throat Singing group. Cluster 2: Exclusively includes papers that study the risks of throat cancer for certain populations. Cluster 3: Isolates the paper “Study on throat singing”. Cluster 4: Includes papers on medical related issues of throat cancer. The “Harmonic overtone singing” and “Overtone singing” papers are also included in this group due to the use of anatomical terms in these papers.

6A.3.4.3 Simple K-means (Cosine) with Three Clusters

Cluster 1: Consists of 8 papers. It includes only Throat Singing papers and as such is a pure cluster.

Cluster 2: Consists of 11 papers with all the Throat Cancer papers, plus “Harmonic overtone singing”.

Cluster 3: Again the “Study on throat singing” paper is isolated in a distinct cluster.

6A.3.4.4 Simple K-means (Cosine) with Four Clusters

Cluster 1: Consists of 8 papers. It includes only Throat Singing papers and as such is a pure cluster.

Cluster 2: Consists of 10 papers with 9 Throat Cancer papers, plus the “Harmonic overtone singing” paper.

Cluster 3: Again the “Study on throat singing” paper is isolated in a distinct cluster.

Cluster 4: One paper on Throat Cancer- a definitive paper on “Cancer - throat or larynx”.

The results in the above experiments consistently show our methods can differentiate between different but closely related topics. The grouping of the papers “Harmonic overtone singing” and “Overtone singing” with the Throat Cancer papers is due to the use of anatomical terms in these papers. Simple K- means with 4 clusters stands out in presenting the most refined groupings.

6A.3.4.5 Cobweb Clustering with 20 Papers

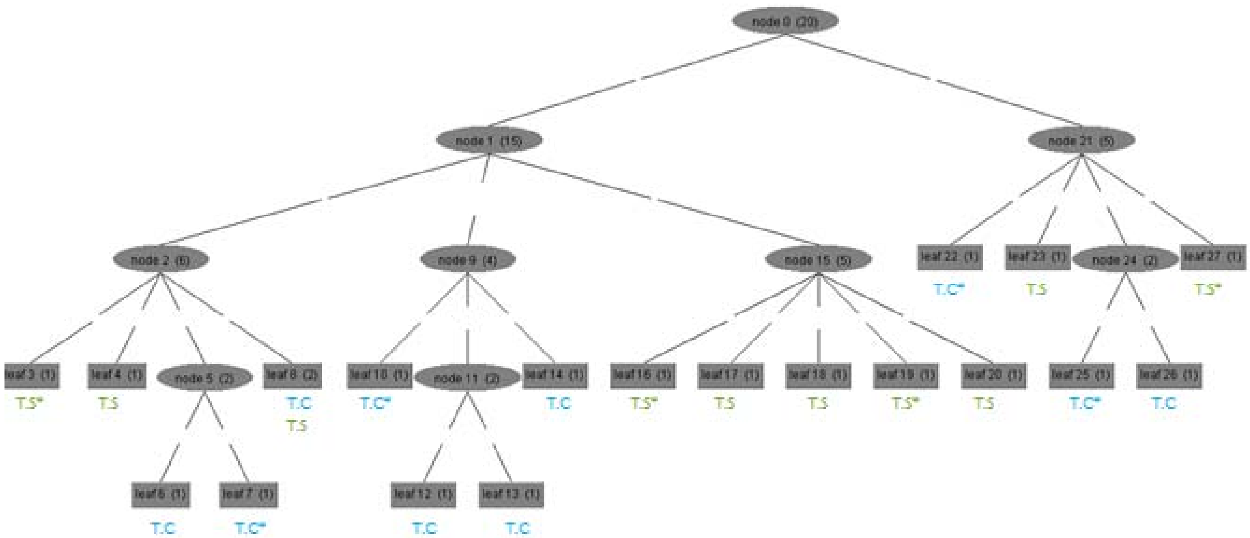


Fig.6A.4.3.5.1: Cobweb clustering on the 20 papers

- Leaf3: Overtone singing
- Leaf4: What is throat singing
- Leaf6: New throat cancer treatment
- Leaf7: Cancer - throat or larynx - Overview
- Leaf8: Harmonic overtone singing, Smoking and cancer of the mouth, pharynx and larynx. These two leaves are actually both children of Node2 (the Cobweb display group leaves sometimes to save horizontal space).
- Leaf10: Ear, nose, and throat cancer : Ultrasound diagnosis of metastasis to cervical lymph nodes
- Leaf12: Oral mucositis in cancer therapy
- Leaf13: Quality of life 5-10 years after primary surgery
- Leaf14: Diet in the etiology of oral and pharyngeal cancer among women from the southern United States
- Leaf16: The throat singers of Tuva
- Leaf17: Inuit throat-games and Siberian throat singing a comparative, historical, and semiological approach
- Leaf18: Mongolian conceptualizations of overtone singing
- Leaf19: Study on throat singing
- Leaf20: A study of the blending of vocal music with the sound field by different singing styles

- Leaf22: 1985 Findings on health promotion and disease prevention
- Leaf23: Tuvan throat singing
- Leaf25: Cancer coverage and tobacco advertising in African-American women's popular magazines
- Leaf26: Perceived risks of certain types of cancer and heart disease among Asian American smokers and non-smokers
- Leaf27: Types of throat singing

If you observe the Cobweb distribution, Node 2 has a mix of Throat Singing and Throat Cancer papers. (*) on the indication denotes that it is a root paper from which the Semantic Signatures were generated. Leaf 3 and 4 are Throat Singing papers. Node 5 is pure with Throat Cancer papers.

Node 9 and its descendents are pure with Throat Cancer papers. They are medical oriented Throat Cancer papers.

Node 15 and its descendents are pure with Throat Singing papers. They are mostly about specific studies and approaches to Throat Singing.

Node 21 is a mixed cluster with both Throat Singing and Throat Cancer, yet its descendent node Node24 is a pure clusters which has non-medical Throat Cancer papers.

Our methods can be used to classify documents by using the root papers (that have known content) as markers for the clusters; papers within a cluster are classified in the genre of the root paper(s) in the cluster. For the hierarchical Cobweb classification, we trace upward from a root paper (indicated by * in the Figure 6A.4.3.5.1) to its nearest internal Node ancestor; all the descendants of this internal Node inherit the genre of the root paper. If we have knowledge of only the root papers, as to what genre they belong to and remaining papers are of unknown categorization, we can use our tool to classify them.

6A.4 Final Conclusion

Subtle differences between two genre/topics are significantly differentiated by SSMinT. Our tool can classify documents with unknown content into their true genres by learning on a few documents. Clearly, there is a subgrouping within the topic, such as a) non-medical versus medical throat cancer papers, and b) cancer risk assessment versus cancer symptoms and treatment. This implies the context in which throat cancer appears has been identified. Our experiments show that our methods are highly effective and sensitive to subtle differences in content. There is a room to conduct further experiments and reproduce such results.

6B: Financial Data Experiment
6B.1 Introduction: What is a 10-K form?

According to [28], a 10-K form is an annual report required by the U.S. Securities and Exchange Commission (SEC) that gives an overall summary of a public company's performance in the market. Although similarly named, the annual report on Form 10-K is different from the "annual report to shareholders" which a company must send to its shareholders when it holds an annual meeting to elect directors. Though, some companies combine the annual report to the shareholders and the Form 10-K into one document. The 10-K includes information such as executive compensation, company history, equity organizational structure, subsidiaries and audited financial statements.

Every annual report contains 4 parts and 15 schedules. They are

PART I

ITEM 1. Description of Business

ITEM 1A. Risk Factor

ITEM 1B. Unresolved Staff Comments

ITEM 2. Description of Properties

ITEM 3. Legal Proceedings

ITEM 4. Submission of Matters to a Vote of Security Holders

PART II

ITEM 5. Market for Registrant's Common Equity, Related Stockholder Matters and Issuer Purchases of Equity Securities

ITEM 6. Selected Financial Data

ITEM 7. Management's Discussion and Analysis of Financial Condition and Results of Operations

ITEM 7A. Quantitative and Qualitative Disclosures About Market Risk

ITEM 8. Financial Statements and Supplementary Data

ITEM 9. Changes in and Disagreements With Accountants on Accounting and Financial Disclosure

ITEM 9A(T). Controls and Procedures

ITEM 9B. Other Information

PART III

ITEM 10. Directors, Executive Officers and Corporate Governance

ITEM 11. Executive Compensation

ITEM 12. Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters

ITEM 13. Certain Relationships and Related Transactions, and Director Independence

ITEM 14. Principal Accounting Fees and Services

PART IV

ITEM 15. Exhibits, Financial Statement Schedules Signatures

6B.2. About the Experiment

To design the experiment, we requested the expertise of Dr. Bonnie Morris, Associate Professor, Department of Business and Economics, WVU [29]. She helped us understand the nature of 10-k files and which part of it would be of our interest.

Out of the 10-K forms, the content that interested us is Item 7 and Item 7A, as these are the management discussion and analysis of the financial conditions. Here, management discusses the operations of the company in detail by usually comparing the current period versus the prior period. These comparisons provide the reader an overview of the operational issues that caused certain increase or decrease in the business.

Since this indicates the performance of each company, we are interested to see if a last 10-K document of a Bankrupt company can predict its closure. Thus, we looked for some comparable companies in the retail industry and started collecting the Item 7 and 7A sections of their 10-K reports. We did this for both Bankrupt and Non-Bankrupt comparable companies in 2009.

No company declares openly that bankruptcy is imminent, and since the format of the 10-k report is more like a boilerplate pattern, they may indicate their bankruptcy subtly in numbers or in text. We were interested to see if SSMinT can predict their bankruptcy from their last 10-K form.

6B.3 Objective

The semantic sensitivity nature of the text can be best ensured in the 10-K reports as they try to showcase their company to be in good shape even though they are not. In such case our objective in devising an experiment was: To predict the bankruptcy of a company with the aid of SSMinT and distinguish these troubled companies from comparable healthy (companies that did not go bankrupt immediately after their 2009 10-K report) .

6B.4 Design of the Experiment

Initially, to see how this experiment would shape up, Dr. Morris [29] helped us select 5 Bankrupt and 5 Non-Bankrupt comparable retail store companies. Out of which we choose 3 each to be the training files.

The training files are:

Bankrupt Companies	Non-Bankrupt companies
Circuit City	Best Buy
Eddie Bauer	Target
Finley Jewelry	Signet

Table 6B.4.1 List of known Bankrupt and Non Bankrupt companies

The remaining files would be Gottschalk's and Samsonite for Bankrupt companies and Coach and Cato for Non-Bankrupt companies.

Training files are exposed to Keyword Tool and Learner Tool to develop the Semantic Signatures. The training and testing sets were given as input to Data Analysis Tool to generate the Document Analysis Matrix. Further, the Data Analysis Tool output was clustered in WEKA.

6B.5 Methodology / Approach in Choosing the Keywords

The training files were given as input to Keyword Tool. These training files are basically text files containing Item 7 and Item 7A content of the 10-K annual reports. Once the file is loaded in to Keyword Tool, the keywords can be chosen. Here we have specially treated the financial jargon phrases in the system. We directed Keyword Tool to treat certain phrases like "account reconciliation" and "comparable stores" as one word. Later when keywords are chosen, we kept in mind not to choose words that would evidently indicate bankruptcy; for example, we ignored words like "increase" or "decrease", etc.

6B.6 Experimental Procedure

- From each training set, 3 different keyword sets were chosen. There are 6 training files in total which yielded 18 keyword sets.
- These 18 keyword sets were given to Learner Tool and the Semantic Signature were generated using the distance measures Euclidean and cosine individually.

- These 18 Semantic Signatures were the input to Data Analysis Tool along with the testing and training files. Training files are sent in as the file markers in the resulting clustering/classification of the Bankrupt / Non-bankrupt companies.
- The Document Analysis Matrix is of the dimensions 10 ×18.
- Later we increased the testing set with an additional 5 Bankrupt and 5 Non-bankrupt companies, making the whole input text files set to be 20 files.

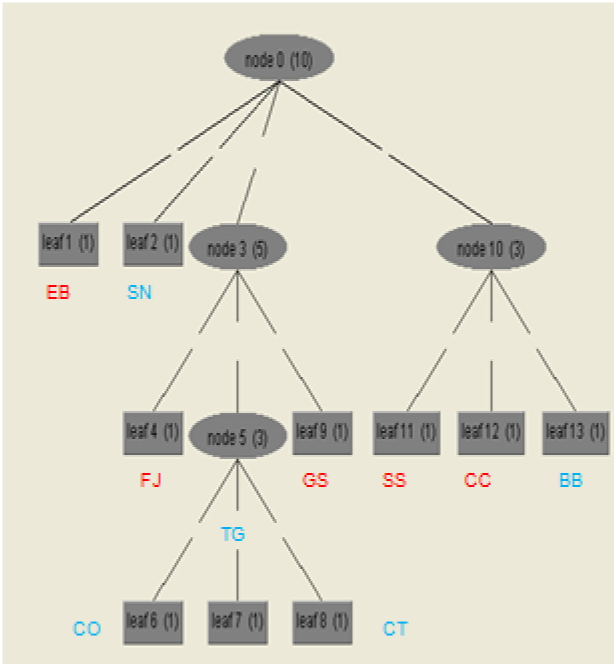
6B.7 Clustering of the DAT Output

We tried K-means clustering on the Document Analysis Matrix with Euclidean and cosine distance measures (while building a ssd, we can select the distance measure for the vectors to cluster). But, the result was all mixed clusters and the interpretation was difficult from the clusters. We were interested to see the hierarchical clustering for this kind of data. For the throat singing and throat cancer experiment, the Cobweb hierarchical clustering gave some good results. We wanted to see if such degree of predictions is possible in this corpus of data.

Cobweb Clustering (Euclidean measure)

Here is the hierarchical breakdown in Figure 6B.7.1. The red highlights are for Bankrupt companies and blue are for Non-bankrupt companies.

Node 5 is a pure cluster with Non-Bankrupt companies. All the remaining internal nodes represent mixed clusters. There is a possibility that these clusters are overlapping with Euclidean distance measures. The boilerplate structure of 10-k reports can be a main reason for such overlapping.



Red – Bankrupt
Blue – Non- Bankrupt

Leaf 1: Eddie Bauer
Leaf 2 : Signet
Leaf 4: Finley Jewelry
Leaf 6: Coach
Leaf 7: Target
Leaf 8: Cato
Leaf 9: Gottschalk
Leaf 11: Samsonite
Leaf 12: Circuit City
Leaf 13: Best Buy

Figure 6B.7.1 Cobweb clustering –Euclidean distance measure

Cobweb clustering (cosine measure)

We wanted to see a sparse distribution of hierarchical clustering with cosine distance measures. We tried the hierarchical clustering on a different Document Analysis Matrix generated using the cosine distance measure Semantic Signatures.

Here is the hierarchical breakdown in Figure 6B.7.2, with two separate nodes that are mostly pure. There is a distinction between two nodes and mostly they are pure except one misgrouped element. *Coach* and *Eddie Bauer* are such misgrouped elements. Nevertheless, the degree of accurate prediction is very high.. Though the structures of the Bankrupt and Non-Bankrupt reports are similar, SSMinT can cluster the corpus into two distinct groups.

We are now ready to increase the testing set, with 5 Bankrupt companies and 5 Non-Bankrupt companies, to see if the package of tools can reproduce this result. Again, the Semantic Signature sets for Euclidean and cosine measures are kept intact.

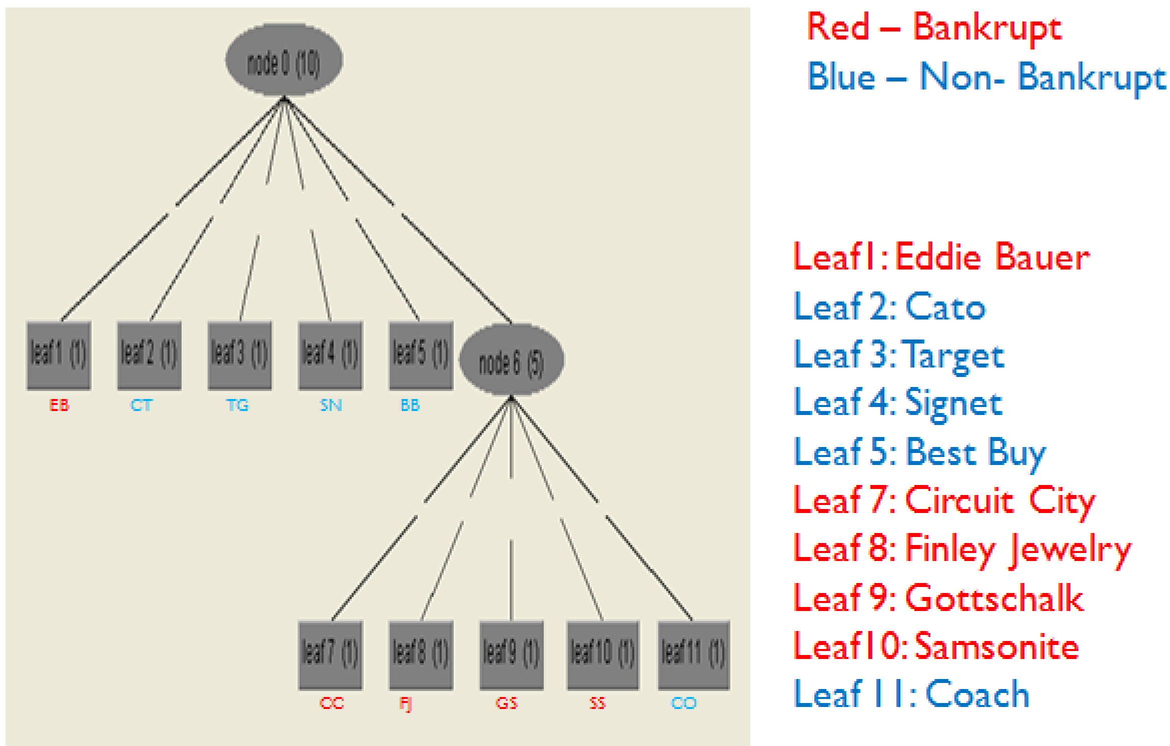


Fig 6B.7.2 Cobweb clustering – cosine distance measure

Cobweb clustering – Euclidean measure for a larger set

Adding 5 bankrupt and 5 non bankrupt companies.

Bankrupt Companies	Non-Bankrupt Companies
Gantos Inc	Advance Auto Parts
Paul Harris stores Inc	RadioShack
Shoe Pavilion	Ann Taylor
Sound Advice	Finish Line
Hartmarx Ross	

Table 6B.7.1 List of unknown Bankrupt and Non-Bankrupt companies

Keeping the Euclidean Semantic Signatures intact we ran all 20 files (including testing and training sets) with the Semantic Signatures in Data Analysis Tool. The new Document Analysis Matrix was subjected to Cobweb clustering in Weka.

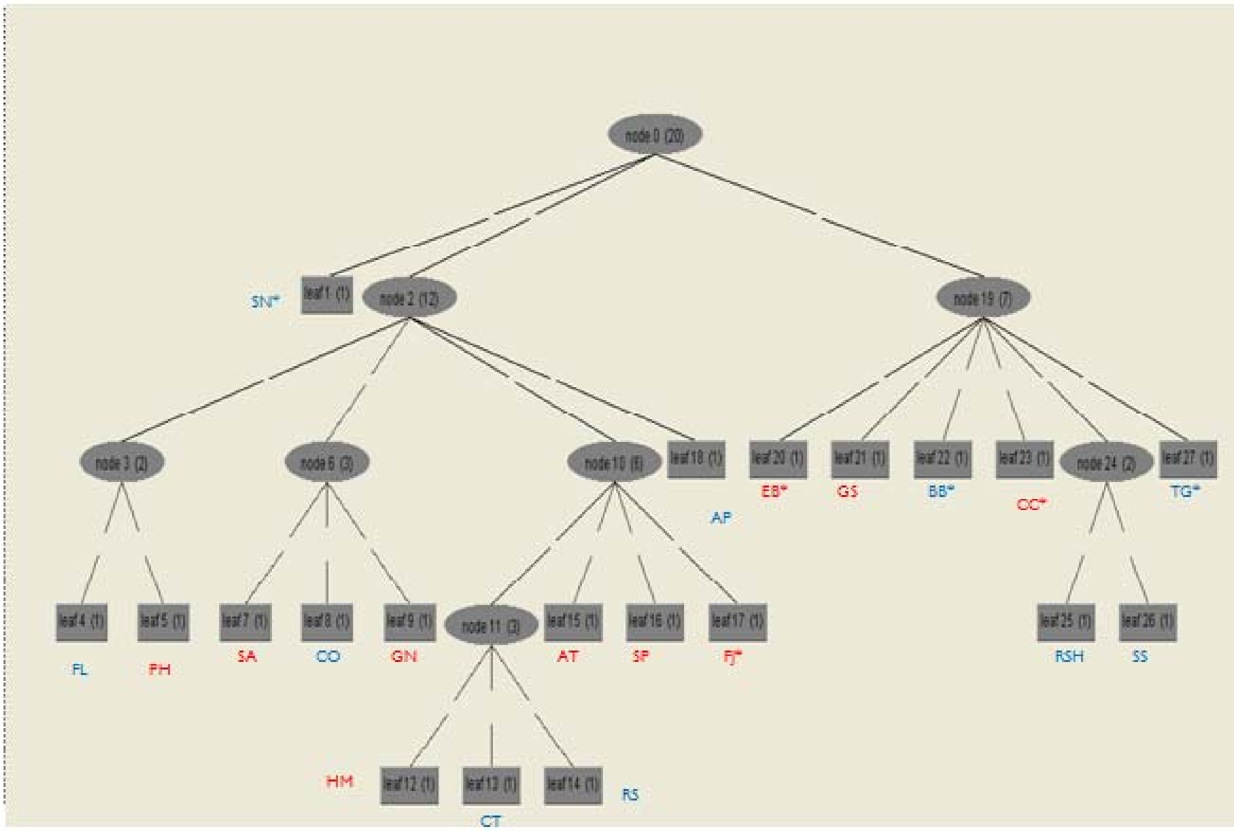


Figure 6B.7.3 Cobweb clustering – Euclidean distance measure (larger set)

Similar to the Euclidean distribut ion with the smaller se t, this hierarchical breakdow n also has a lot of overlapping. W e cannot conclude any inform ation out of such clustering output. There are certain pure nodes and e qually m ixed nodes. To further analyz e the output, we took the cosine distribution under consideration.

Cobweb clustering – Cosine measure for a larger set

The breakdown of the Cobweb clustering with th e cosine distan ce measure is sho wn in Figure 6B.7.4. The hierarchical clustering is neat and very impressive. The degree of accurate prediction also is high.

If we are blind folded from the knowledge of the category of testing files, the training files act as file markers and prediction is possible. For exam ple, observe Node 1, here the category of Leaf7 which has Target is known as a Non-Bankrupt com pany and thus, we can m ove up to the parent node and predict that all the leaves under Node 1 are "Non-Bankrupt" companies.

Observe Node 16, here we have a pure node and its descendents. There are two training files which are Bankrupt companies among the descendants of Node 16, so we can predict that Node 16 is a "Bankrupt" node (i.e., all the descendants of Node 16 are Bankrupt companies).

Node 8 is an “indeterminant” node as it has both Bankrupt and Non-Bankrupt training files under it and we cannot determine the category of the files present under this node

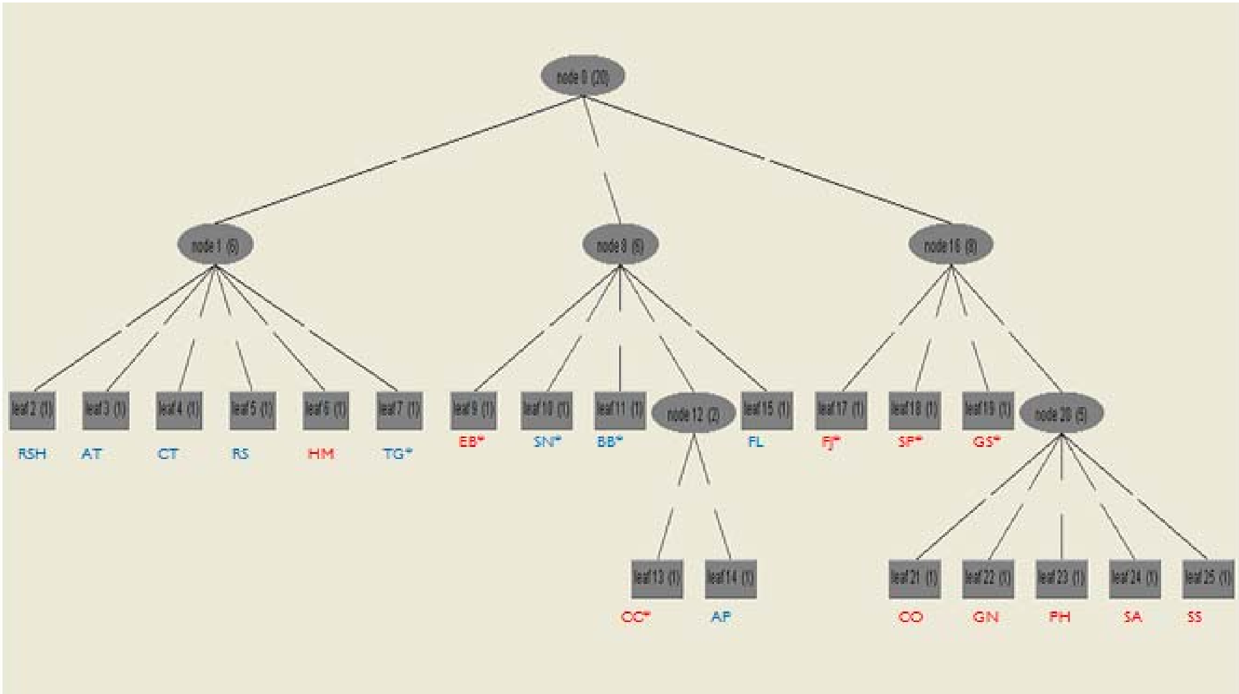


Figure 6B.7.4 Cobweb clustering- cosine distance measure (larger set)

6B.8 Final Conclusion

SSMinT can differentiate the Bankrupt versus Non-Bankrupt companies on one condition: the Semantic Signatures must be chosen intelligently. That is an expert is required to choose keywords and identify important phases in such a way as to capture the subtle differences in 10-K reporting between companies that will soon file for Bankruptcy and healthy companies.. An expert is required to choose Semantic Signatures that best model the nuances of the language used. This paves the way towards automating understanding the semantic sensitive nature of the English language. Further experiments are required to show that results are reproducible and show the effectiveness of our methods on larger data sets.

7: Future Work

In this thesis, a novel text mining tool was presented which is based on capturing the content in a text document. Core modules in the SSMinT package like keyword selection, Semantic Signature development were introduced in detail. The process of the whole framework was also defined. A series of experiments with different corpora demonstrated that the proposed method is feasible and effective in the text mining.

Future work with the SSMinT package will be to reduce the burden on the analyst or the expert who uses the tools. The knowledge about the corpus is currently a requirement when it comes to selection of apt keyword sets. But, if this burden on the analyst can be automated, the tools will be powerful in the hands of even analysts who are non-experts in the input corpora.

To waive the analyst's intervention to a certain level, we have proposed the process of automating the whole process of keyword selection, Semantic Signature development and the pruning of Semantic Signatures. After employing certain algorithms for decision making, we can certainly prune the Semantic Signatures, and once automated, the tool will have a great scope of reaching the common audience.

Pruning the Semantic Signatures to include only those that capture significant attributes of the target content is an important functionality for the SSMinT tool. The curse of high dimensionality is that it limits the system to not present proper results by including unnecessary dimensions which cause noise in the system. Once the Semantic Signatures are evaluated and learning takes place on them, we can prune the Semantic Signatures that do not aid in the system's performance. By doing so, we are removing unnecessary noise within the system.

The semantic sensitivity experiments were explored with only one type of hierarchical clustering available in Weka, the Cobweb clustering. In the future, we might want to expose the output of Data Analysis Tool to various types of hierarchical clustering techniques.

Language independence is another area of research, though SSMinT is totally independent of any language except the stemming plug-in. We are dealing with the issues of proper display of the Unicode characters in the text point back function. Next, we will test the full functionality on foreign language such as Hindi and Telugu.

Data visualization and software enhancements are required for the current SSMinT package. Improvement in the visualizing the document vectors and their clusters is very desirable.

We will continue to investigate the capability of Semantic Signatures to embody and quantify emotive shift in the text data. This most likely will utilize phrase keywords and require extending our Semantic Signatures to include intensity ranking of meta-words. Special handling of expletives and hate words may also be of value.

The new framework for text mining presented here will open a wide range of applications and possibilities in text mining and the above exciting challenges will be addressed in the future work.

References

[1] H. A. Do Prado and E. Ferneda. (2007), *Emerging Technologies of Text Mining: Techniques and Applications* .

[2] Q. Wu, E. Fuller and C. Zhang. (2010), "Gra ph m odel for pattern recognition in text," in *Studies in Computational Intelligence* (V.288 ed.), I. Ting, H. -. Wu and T. -. Ho, Eds.

[3] Q. Wu, E. Fuller and C. Zh ang. (2009), Text document classification and pattern recognition. *Social Network Analysis and Mining, International Conference on Advances in 0*pp. 405-410.

[4] J. Franke, G. Nakhaei zadeh and I. Renz. (2003), *Text Mining, Theoretical Aspects and Applications* .

[5] A. Stavrianou, P. Andritsos an d N. Nicoloyannis. (2007), Over view and se mantic issues of text mining. *SIGMOD Record 36(3),* pp. 23-34.

[6] W ikipedia. (2010), Text m ining --- wikipedia, the free encyclopedia. Available: http://en.wikipedia.org/w/index.php?title=Text_mining&oldid=394838873.

[7] R. Feldm an and J. Sanger. (2006), *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data* .

[8] S. Weiss, N. Indurkhya, T. Zh ang and F. Dam erau. (2004, October). *Text Mining: Predictive Methods for Analyzing Unstructured Information* Available: http://www.worldcat.org/isbn/0387954333.

[9] C. D. Manning, P. Raghavan and H. Schtze. (2008), *Introduction to Information Retrieval* .

[10] A. Amir, Y. Aumann, R. Feldman and M. Fresko. (2005), Nove mber). Maximal association rules: A tool for m ining associations in text. *J.Intell.Inf.Syst. 25(3),* pp. 333-345. Available: http://portal.acm.org/citation.cfm?id=1107361.1107392.

[11] W. R. Hersh. (2005), Evalua tion of biom edical text-mining systems: Lessons learned from information retr ieval. *Briefings in Bioinformatics 6(4),* pp. 344-356. Available: http://dblp.uni-trier.de/db/journals/bib/bib6.html#Hersh05.

[12] H. Ya ng and C. Lee. (2005), Autom atic category them e identification and hierarchy generation for chinese text categorization. *J Intell Inform Syst 25(1),* pp. 47-67. Available: http://dx.doi.org/10.1007/s10844-005-0859-6.

[13] J. Turmo, A. Ageno and Catal`a Neus. (2006, July). Adaptive in formation extraction. *ACM Comput.Surv. 38(2),* Available: http://doi.acm.org/10.1145/1132956.1132957.

[14] M. Saravanan, P. C. R. Raj and S. Ra man. (2003), SUMMARIZATION AND CATEGORIZATION OF TE XT DAT A IN HIGH-L EVEL DATA CLE ANING F OR

INFORMATION RETRIEVAL. *Applied Artificial Intelligence: An International Journal 17(5),* pp. 461. Available: http://www.informaworld.com/10.1080/713827177.

[15] P. Srinivasan. Text m ining: Generating hypotheses from MEDLINE. *J. Am. Soc. Inf. Sci. Technol. 55*pp. 396-413.

[16] D. va n Heijst, R. Potharst and M. van W ezel. (2008, March). A support system for predicting eBay end prices. *Decis. Support Syst. 44(4),* pp. 970-982. Available: http://dx.doi.org/10.1016/j.dss.2007.11.004.

[17] Richard S. Segall, Qingyu Zhang and Me i Cao, "Web-Based Text Mining of Hotel Customer Comments Using SAS® Text Miner and Megaputer Polyanalyst®," .

[18] W ikipedia. (201 0, Stemmi ng --- wikipedi a, th e free encycloped ia. Available : http://en.wikipedia.org/wiki/Stemming.

[19] Lancaster University, "What is Porter Stemming?" .

[20] W ikipedia. (2010), Stop words --- w ikipedia, th e free encyclopedia. Available: http://en.wikipedia.org/wiki/Stop_words.

[21] H. Arimura, J. Abe, R. Fujino, H. Sakam oto, S. Shimozono, S. Arikawa and S. Shim ozono. Text data m ining: Discovery of im portant keywords in the cyberspace . Presented at 2000 Kyoto International Conference on Di gital Libraries : Research and Practice. Availab le: http://dx.doi.org/10.1109/DLRP.2000.942178.

[22] U. K. Para, "Com puter-aided Sem antic Signatu re Iden tification and Docum ent Classification via Semantic Signatures," 2010.

[23] M. Hall, E. Frank, G. Hol mes, B. Pfahri nger, P. Reutem ann and I. H. W itten. (2009), The WEKA data m ining software: An update. *SIGKDD Explor.Newsl. 11(1),* pp. 10-18. Available: http://dx.doi.org/10.1145/1656274.1656278.

[24] J. A. Goldsm ith, D. Higgins and S. Soglasnova. Autom atic language-specific stemming in information retrieval. P resented at In Cross- Language Inform ation Retrieval and Evaluation: Proceedings of the CLEF 2000 Workshop.

[25] tartarus.org. (2006), The porter stemming algorithm.

[26] T. Rose, M. Stevenson and M. W hitehead. The reuters corpus volum e 1 - from yesterday's news to tomorrow's lan guage resources. Presented at In Proceedings o f the Third Intern ational Conference on Language Resources and Evaluation.

[27] S. Sve dman. (1970, Apr.). Sem antic sensitiv ity and reading ach ievement. *The Reading Teacher 23(7, Primary Reading),* pp. pp. 640-646, 648. Available: http://www.jstor.org/stable/20196388.

[28] W ikipedia. (2010), Form 10-K --- wiki pedia, th e free encyclopedia. Available: http://en.wikipedia.org/wiki/Form_10-K.

[29] B. Morris, "Bankrupt and Non- Bankrupt companies," vol. Personal communication, 2010.